I lettori che desiderano informarsi sui libri e sull'insieme delle attività della Società editrice il Mulino possono consultare il sito Internet:
www.mulino.it

ISBN 978-88-15-13733-3

Copyright © 2004 by Società editrice il Mulino, Bologna. Nuova edizione 2010. Tutti i diritti sono riservati. Nessuna parte di questa pubblicazione può essere fotocopiata, riprodotta, archiviata, memorizzata o trasmessa in qualsiasi forma o mezzo – elettronico, meccanico, reprografico, digitale – se non nei termini previsti dalla legge che tutela il Diritto d'Autore. Per altre informazioni si veda il sito **www.mulino.it/edizioni/fotocopie**

Indice

Premessa 13

1 Statistica e metodologia della ricerca 15
1.1 Introduzione . . . 15
1.2 Evoluzione storica della Statistica . . . 15
1.3 La collocazione scientifica della Statistica . . . 17
1.4 Alcuni paradigmi della ricerca statistica . . . 18
1.5 Problemi reali e problemi statistici . . . 19
1.6 Fasi di una analisi statistica . . . 21
1.7 Il percorso logico di questo volume . . . 23
1.8 Considerazioni finali . . . 24

Ambiente R 25
R1.1 Cosa è l'ambiente statistico R . . . 25
R1.2 Come si attiva una sessione di R . . . 26
R1.3 Richiesta di aiuto in R . . . 28
R1.4 Nozioni di base per lavorare in R . . . 28
R1.5 Lavorare in R . . . 30
R1.6 I data set utilizzati per le analisi . . . 31
R1.7 Riferimenti bibliografici per R . . . 32

I Analisi esplorativa delle informazioni 35

2 Rilevazioni statistiche 37
2.1 Introduzione . . . 37
2.2 Elementi di una rilevazione statistica . . . 37
2.3 Tipologia delle informazioni statistiche . . . 42
2.4 Operazioni statistiche elementari . . . 46

2.5 Rapporti statistici 47
2.6 Presentazione delle rilevazioni statistiche 50
2.7 Considerazioni finali 54

Ambiente R **57**
R2.1 Gestione delle variabili in R 57
R2.2 Natura delle variabili in R 57
R2.3 Gli oggetti di R 59
R2.4 Analisi di un vettore 59
R2.5 Analisi di una variabile quantitativa 63
R2.6 Matrici 65
R2.7 *Data frame* 67
R2.8 Operazioni di input e output in R 71
R2.9 Rappresentazioni grafiche 73
R2.10 Rappresentazioni di variabili qualitative 77
R2.11 Rappresentazioni di variabili quantitative 80
R2.12 Cenni alle liste 82
R2.13 Rappresentazioni di serie storiche 83

3 Distribuzioni di frequenza **89**
3.1 Introduzione 89
3.2 Rappresentazioni tabellari e grafiche 89
3.3 Distribuzioni per variabili discrete 94
3.4 Distribuzioni per variabili continue 95
3.5 La funzione di ripartizione empirica 98
3.6 Indicatori sintetici delle variabili statistiche 102

Ambiente R **107**
R3.1 Distribuzioni di frequenza e natura delle variabili 107
R3.2 Variabili qualitative 108
R3.3 Variabili quantitative discrete 109
R3.4 Variabili quantitative continue 112
R3.5 Funzione di ripartizione empirica 114
R3.6 Normalizzazione di una variabile 115

4 Indici statistici di posizione **117**
4.1 Introduzione 117
4.2 Il concetto di media 118
4.3 La media aritmetica 119
4.4 Proprietà e difetti della media aritmetica 123
4.5 La media aritmetica ponderata 126
4.6 Moda e mediana 132
4.7 Quantili di una distribuzione di frequenza 140
4.8 Altri indici di posizione 142
4.9 Considerazioni finali 145

Ambiente R **147**
R4.1 Media aritmetica 147

R4.2 Moda, mediana e quantili 149
R4.3 Altri indici di posizione 150

5 Indici statistici di variabilità **153**
5.1 Introduzione 153
5.2 Variabilità rispetto ad un centro 154
5.3 Variabilità delle modalità ordinate 159
5.4 Mutua variabilità 161
5.5 Concetto e misura della concentrazione 164
5.6 Indici di eterogeneità 172
5.7 Considerazioni finali 175

Ambiente R **177**
R5.1 Variabilità rispetto ad un centro 177
R5.2 Variabilità dalle modalità ordinate 178
R5.3 Mutua variabilità 179
R5.4 Curva di Lorenz 179
R5.5 Indici di eterogeneità 181

6 Forma di una distribuzione di frequenza **183**
6.1 Introduzione 183
6.2 Asimmetria 183
6.3 Curtosi di una distribuzione di frequenza 187
6.4 Forma e istogramma perequato 189
6.5 Metodi esplorativi e distribuzioni di frequenza 190
6.6 Considerazioni finali 191

Ambiente R **193**
R6.1 Asimmetria e curtosi di una distribuzione 193
R6.2 Istogramma perequato 194
R6.3 Box-plot 195
R6.4 Trasformazioni delle variabili 200
R6.5 Operazioni ripetute in R 202

7 Distribuzioni statistiche multiple **205**
7.1 Introduzione 205
7.2 Distribuzioni doppie di frequenza 205
7.3 Distribuzioni condizionate e indipendenza 213
7.4 Una misura di connessione tra variabili doppie 216
7.5 Correlazione tra variabili statistiche 218

Ambiente R **229**
R7.1 Introduzione 229
R7.2 Distribuzioni doppie di frequenza 229
R7.3 Rappresentazioni grafiche di variabili doppie 231
R7.4 Misure di connessione 236
R7.5 Correlazione tra variabili statistiche 237

II Teoria della probabilità 239

8 Calcolo delle probabilità 241
8.1 Introduzione . . . 241
8.2 Concetto e misura della probabilità . . . 242
8.3 Caratteristiche degli esperimenti probabilistici . . . 243
8.4 Concezioni alternative . . . 246
8.5 Assiomatizzazione del Calcolo delle probabilità . . . 249
8.6 Concetti primitivi . . . 251
8.7 Formalizzazione delle prove . . . 253
8.8 Formalizzazione degli eventi . . . 255
8.9 Postulati del Calcolo delle probabilità . . . 264
8.10 Principali teoremi del Calcolo delle probabilità . . . 265
8.11 Probabilità condizionata e indipendenza . . . 267
8.12 La misura della probabilità . . . 270
8.13 Il Teorema di Bayes . . . 280
8.14 Alcune esemplificazioni . . . 286
8.15 Considerazioni finali . . . 292

Ambiente R 297
R8.1 Estrazione di palline da un'urna . . . 297
R8.2 Problema del compleanno . . . 298
R8.3 Un problema di Calcolo delle probabilità . . . 298

9 Variabili casuali 303
9.1 Introduzione . . . 303
9.2 Definizione di variabile casuale . . . 304
9.3 Variabili casuali discrete . . . 307
9.4 Variabili casuali continue . . . 310
9.5 La funzione di ripartizione di una variabile casuale . . . 312
9.6 Valori medi di una variabile casuale . . . 318
9.7 Variabili casuali multiple . . . 325
9.8 Variabili casuali doppie . . . 326
9.9 Covarianza e correlazione tra variabili casuali . . . 333

Ambiente R 337
R9.1 Rappresentazione di una variabile casuale discreta . . . 337
R9.2 Rappresentazione di una variabile casuale continua . . . 338

10 Principali modelli probabilistici 339
10.1 Introduzione . . . 339
10.2 La variabile casuale Uniforme continua . . . 340
10.3 La variabile casuale Normale . . . 342
10.4 Variabili casuali connesse alla variabile casuale Normale . . . 349
10.5 La variabile casuale Uniforme discreta . . . 351
10.6 La variabile casuale di Bernoulli . . . 353
10.7 La variabile casuale Binomiale . . . 354
10.8 La variabile casuale di Poisson . . . 358

Ambiente R 363
R10.1 Famiglie parametriche di variabili casuali . . . 363
R10.2 Variabile casuale Uniforme continua . . . 364
R10.3 Variabile casuale Normale (Gaussiana) . . . 365
R10.4 Variabili casuali Chi-quadrato, t ed F . . . 368
R10.5 Variabile casuale Uniforme discreta . . . 369
R10.6 Variabile casuale Binomiale . . . 369
R10.7 Variabile casuale di Bernoulli . . . 371
R10.8 Variabile casuale di Poisson . . . 373

11 Teoremi asintotici 375
11.1 Introduzione . . . 375
11.2 Leggi dei grandi numeri . . . 376
11.3 Teorema Limite Centrale . . . 378
11.4 Altri teoremi asintotici . . . 385
11.5 Elementi di teoria della simulazione . . . 385

Ambiente R 391
R11.1 Teorema Limite Centrale . . . 391
R11.2 Un esperimento di simulazione . . . 393

III Inferenza statistica 397

12 Campioni casuali e distribuzioni campionarie 399
12.1 Introduzione . . . 399
12.2 Principi dell'inferenza statistica . . . 400
12.3 Procedure inferenziali . . . 404
12.4 Campioni casuali e distribuzioni campionarie . . . 406
12.5 Media e varianza campionarie . . . 410
12.6 Altre distribuzioni campionarie notevoli . . . 413
12.7 Simulazione di una distribuzione campionaria . . . 414
12.8 La funzione di verosimiglianza . . . 417

Ambiente R 425
R12.1 Un esempio di distribuzione campionaria . . . 425
R12.2 La funzione di verosimiglianza . . . 427

13 Teoria della stima 431
13.1 Introduzione . . . 431
13.2 Stimatori e stime di un parametro . . . 431
13.3 Sufficienza di uno stimatore . . . 436
13.4 Proprietà finite di uno stimatore . . . 439
13.5 Proprietà asintotiche di uno stimatore . . . 448
13.6 Principi generali per la stima di un parametro . . . 452
13.7 Metodi di costruzione degli stimatori . . . 455
13.8 Metodo della massima verosimiglianza . . . 462

Ambiente R 469
R13.1 Metodo della massima verosimiglianza . . . 469
R13.2 Metodo del minimo Chi-quadrato . . . 470

14 Test delle ipotesi statistiche 473
14.1 Introduzione . . . 473
14.2 Caratteristiche fondamentali di un test . . . 475
14.3 Logica del test delle ipotesi . . . 479
14.4 Struttura probabilistica del test . . . 483
14.5 Lemma di Neyman e Pearson . . . 489
14.6 Procedura operativa per un test delle ipotesi . . . 495
14.7 Test sui parametri di una v.c. Normale . . . 496
14.8 Test sulla differenza tra valori medi di v.c. Normali . . . 501
14.9 Test sul rapporto tra le varianze di v.c. Normali . . . 504
14.10 Test sul confronto tra proporzioni . . . 506
14.11 Test sulla bontà di adattamento . . . 509
14.12 Test sulla indipendenza . . . 515
14.13 Considerazioni finali . . . 522

Ambiente R 525
R14.1 Un esempio di applicazione del test . . . 525
R14.2 Test sui valori medi di v.c. Normali . . . 526
R14.3 Test sul rapporto tra varianze . . . 529
R14.4 Test sul confronto tra proporzioni . . . 530
R14.5 Test sulla bontà di adattamento . . . 530
R14.6 Test sulla indipendenza . . . 531

15 Intervalli di confidenza 533
15.1 Introduzione . . . 533
15.2 Intervalli casuali e di confidenza . . . 534
15.3 Intervallo di confidenza per il valore medio . . . 534
15.4 Intervalli di confidenza parametrici . . . 541
15.5 Intervalli di confidenza asintotici . . . 545
15.6 Intervalli di confidenza e test delle ipotesi . . . 547
15.7 Considerazioni finali . . . 549

Ambiente R 551
R15.1 Generazione di intervalli di confidenza . . . 551
R15.2 Intervalli di verosimiglianza . . . 552

16 Elementi di campionamento statistico 555
16.1 Introduzione . . . 555
16.2 Elementi di un piano campionario . . . 556
16.3 Principali piani di campionamento . . . 559
16.4 Determinazione della numerosità campionaria . . . 566
16.5 Campionamento con risposta casualizzata . . . 570
16.6 Ulteriori piani di campionamento . . . 573
16.7 Considerazioni finali . . . 576

Ambiente R 581
R16.1 Campionamento con e senza ripetizione . . . 581

IV Introduzione ai modelli statistici 583

17 Modelli statistici 585
17.1 Introduzione . . . 585
17.2 Modelli e procedure inferenziali . . . 585
17.3 Costruzione di un modello statistico . . . 588
17.4 Classificazione dei modelli . . . 593
17.5 Considerazioni finali . . . 599

Ambiente R 601
R17.1 Modelli statistici in R . . . 601
R17.2 Utilizzo della funzione lm() . . . 601
R17.3 Output della funzione lm() . . . 602

18 Il modello di regressione lineare 605
18.1 Introduzione . . . 605
18.2 Modello di regressione semplice . . . 606
18.3 Aspetti inferenziali del modello . . . 609
18.4 Specificazione del modello . . . 611
18.5 La stima dei parametri . . . 614
18.6 Proprietà degli stimatori . . . 621
18.7 Verifica del modello stimato . . . 624
18.8 Indice di determinazione multipla . . . 628
18.9 Utilizzazioni del modello di regressione semplice . . . 635
18.10 Considerazioni finali . . . 637

Ambiente R 639
R18.1 Modello di regressione lineare semplice . . . 639
R18.2 Un'applicazione del modello di regressione . . . 641
R18.3 Modello di regressione con variabili indicatrici . . . 644
R18.4 Discussione di un caso reale . . . 645

Tavole statistiche 653

Indice analitico 663

Indice delle funzioni in R 671

Premessa

La Statistica è una scienza onnipresente nella ricerca scientifica e nelle analisi sperimentali e viene qui proposta come un momento fondamentale della conoscenza umana, che si fonda sull'esperienza e sulle evidenze empiriche. In tale prospettiva, senza indugiare su formule e sviluppi matematici, questo volume insiste sulla logica del metodo statistico e sulle sue connessioni con le vicende quotidiane, alle quali la Statistica apporta un contributo significativo per prendere decisioni razionali e coerenti.

Tale prospettiva è resa attuale dalla riforma universitaria che da un lato ha moltiplicato gli spazi di insegnamento della Statistica e dall'altro l'ha relegata in ambiti ristretti, assieme a discipline solo apparentemente simili (Matematica, Informatica, etc.). Invece, questo libro punta a sottolineare il ruolo proprio della Statistica e la sua valenza scientifica in ordine alla conoscenza finalizzata alla comprensione e all'azione.

A quanti si avvicinano alla Statistica per la prima, e spesso unica volta, è fondamentale comunicare quel percorso unitario che costituisce l'essenza stessa della nostra disciplina: l'analisi esplorativa delle informazioni nasce per comprendere ed interpretare probabilisticamente il mondo e si finalizza alle decisioni, che sono rese operative mediante modelli efficienti, ma sempre perfettibili.

Il testo ripercorre il precedente volume dell'Autore (*Statistica*, Bologna, Il Mulino, 2000), ben più ampio e formalizzato, assumendone la filosofia, la sequenzialità e molti aspetti didattici; tuttavia, esso è stato completamente riscritto per renderlo più semplice ed immediato per un pubblico universitario che non necessariamente sostiene altri esami di Statistica, in aggiunta a quello di base.

Il volume si accompagna con insiemi di dati reali (i cosiddetti *data set*) reperibili sul sito **www.mulino.it/aulaweb** e che sono utilizzati nella quasi totalità degli esempi e delle esercitazioni.

Peraltro, le dimensioni delle informazioni oggi disponibili e la necessità di gestire in modo celere ed automatico le molteplici elaborazioni grafiche e statistiche hanno imposto l'esigenza di avvalersi di strumenti informatici che fossero efficienti e possibilmente disponibili per tutti. In questa filosofia, un posto di rilievo per

gli statistici merita R, un ambiente di programmazione disponibile gratuitamente, che consente sia semplici calcoli che complessi studi su enormi masse di dati. Soprattutto, esso sta divenendo lo standard di riferimento per l'insegnamento, la ricerca e la diffusione dei metodi statistici, in tutto il mondo, per cui appare essenziale poterlo inserire all'interno di un testo universitario di Statistica allo scopo di fornire alle nuove generazioni una professionalità che diventerà sempre più richiesta nel prossimo futuro.

Tale obiettivo è stato perseguito in questo volume aggiungendo, alla fine di ciascun Capitolo, un "Ambiente R" ovvero una lista di istruzioni e commenti sulle analisi del libro sviluppate mediante questo linguaggio reso disponibile per ogni lettore. Come accade per ogni nuova avventura, l'inizio può sembrare faticoso ed incerto ma, progressivamente, chi è tenace e paziente raccoglie molti frutti e, poi, non abbandonerà più un software che è efficiente, veloce, accurato, sempre aggiornato, adattabile ed anche gratuito.

Per una verifica personale dello studio, sempre su **Aulaweb** vi è la possibilità di esercitarsi, capitolo per capitolo, mediante *test di autovalutazione*, seguendo le istruzioni ivi allegate. In particolare, è possibile eseguire i test, conoscere le risposte corrette, ricevere punteggi, memorizzare i risultati, costruire proprie collezioni di quesiti, personalizzando lo studio e ricevendo sempre una valutazione dei risultati conseguiti.

Il testo ha beneficiato della lettura critica e costruttiva di molti colleghi e studenti che ne hanno migliorato la valenza didattica. Anche per questo, sono sempre graditi quesiti, critiche, suggerimenti, correzioni e segnalazioni scrivendo all'Autore via E-mail: *domenico.piccolo@unina.it*.

Il privilegio di insegnare e fare ricerca assieme a Marcella Corduas, Francesca Di Iorio, Carmela Cappelli, Silvia Polettini e Maria Iannario produce risultati scientifici e didattici di notevole valore e questo testo ne è testimonianza concreta: a loro va il mio ringraziamento sincero accomunato alla riconoscenza sostanziale per la vicinanza umana e la collaborazione editoriale di Stefania Capecchi.

Dedico ai miei figli Marco e Francesco questa fatica resa possibile da una famiglia eccezionale che continua a sostenermi con infinita pazienza ed incredibile affetto; è solo l'amore di mia moglie Emilia che rende tuttora piacevole e possibile questo lavoro.

Napoli, 22 gennaio 2010

DOMENICO PICCOLO

Avvertenza: La numerazione di Capitoli, paragrafi, Tabelle e Figure è quella usuale; si premette la lettera R quando sono inseriti nell'Ambiente R. Tuttavia, all'interno di tale ambiente, ove non si produca confusione, la numerazione riprende in modo autonomo.

Capitolo primo

Statistica e metodologia della ricerca

1.1 Introduzione

La *Statistica* è una scienza relativamente giovane il cui contenuto non è tuttora manifesto in modo corretto, perché tale disciplina viene spesso confusa con le *statistiche*: dati, tabelle, grafici, indici, medie. In effetti, essa è presente in tutte le scienze e rappresenta uno strumento essenziale per la scoperta di leggi e relazioni tra fenomeni.

In tutte le situazioni nelle quali occorre assumere decisioni in condizioni di incertezza, la Statistica è un momento importante ed ineliminabile per la ricerca scientifica, le applicazioni della tecnologia, il monitoraggio e la valutazione dei sistemi complessi, la pianificazione economica e l'azione politica. In questi ambiti, la Statistica forma ed aiuta il momento conoscitivo dell'Uomo, fornendogli strumenti formali rigorosi e corroborati dalle osservazioni.

1.2 Evoluzione storica della Statistica

L'etimologia della parola «Statistica» deriva dal vocabolo italiano «Stato» e fa riferimento, nella quasi totalità dei linguaggi europei, alla constatazione per cui le prime informazioni su fenomeni reali (per motivi militari, religiosi, economici, sociali, sanitari, etc.) sono state raccolte ed organizzate ad opera degli organismi statali che ne erano anche i principali utilizzatori. Infatti, la prima apparizione del vocabolo «statistica» in questa accezione sembra essere quella dell'italiano Ghislini che, nel 1589, indica la Statistica come «*descrizione delle qualità che caratterizzano e degli elementi che compongono uno Stato*».

Tuttavia, l'evoluzione storica della Statistica nasconde due anime che si ritrovano ancora oggi nella didattica, nella ricerca, nel pensare comune dei non specialistici e, quindi, nel linguaggio dei mass-media. Ciò provoca qualche confusione in gran parte dell'opinione pubblica e genera spesso una certa diffidenza verso le deduzioni derivate dall'analisi di dati statistici. È fondamentale, quindi, per una corretta comprensione del nostro percorso, evidenziare subito presupposti, analogie e connessioni tra questi due approcci alla Statistica.

• La prima anima della Statistica è connaturata all'*Homo sapiens* che prende coscienza del mondo e dei suoi simili e che, nella lotta per la sopravvivenza, elabora comportamenti ottimali per nutrirsi, difendersi e riprodursi. Tali strategie alternano sconfitte e successi ma, grazie ad approssimazioni successive, permettono il diffondersi di acquisizioni e convinzioni, usi e costumi e, alla fine, di conoscenza, scienza e cultura. Quando tali convinzioni si codificano nella vita del clan, della tribù, di un popolo o di una nazione, nascono le regole di comportamento e, quindi, la legge. Questa esigenza conoscitiva cresce con lo sviluppo delle relazioni tra popoli ed assume infine un connotato obbligatorio e vincolante sino ad essere codificata dalla religione o dall'autorità centrale.

Con la nascita dei grandi Stati europei, si attribuisce all'analisi statistica dei fenomeni collettivi un interesse pubblico che spinge progressivamente le nazioni occidentali a dotarsi di Istituti «centrali» di Statistica, deputati per legge alla raccolta, organizzazione e diffusione di dati sulla popolazione, sulle abitazioni, sulle risorse economiche e – con il passare del tempo – su tutti gli aspetti rilevanti della vita di una nazione, di una Comunità di stati (Unione Europea) o dell'intero pianeta (Nazioni Unite).

Oggi, gli organismi pubblici che istituzionalmente raccolgono e diffondono informazioni statistiche sono innumerevoli ed agiscono secondo una gerarchia di competenze che individua nell'Ente locale (per esempio, il Comune) la sede prioritaria di raccolta del dato elementare, mentre la verifica, l'aggregazione e la pubblicazione sono di competenza dell'Ente centrale (per l'Italia, è l'ISTAT: Istituto Nazionale di Statistica). L'armonizzazione tra le varie fonti di dati e la regolarità delle molteplici rilevazioni sono state inserite, nel nostro Paese, in un sistema articolato definito SISTAN (=Sistema Statistico Nazionale) cui afferiscono le istituzioni pubbliche e private deputate alla raccolta, alla elaborazione e alla diffusione di dati di interesse per la collettività.

Tale quadro si completa con EUROSTAT, organismo della Unione Europea che ha il compito di armonizzare la raccolta, l'analisi e la presentazione dei dati ufficiali sulle nazioni aderenti allo scopo di consentire confronti immediati e verifiche semplici della comune crescita economica, sociale e culturale.

• La seconda anima della Statistica (che solo da pochi secoli ha trovato una formalizzazione compiuta) nasce da una constatazione differente: di fronte alla realtà che muta, vi sono risultati che meritano più fiducia di altri perché si ripetono con maggiore regolarità.

Ciò viene percepito anzitutto in rapporto al clima e all'alternanza delle stagioni ma riguarda anche i raccolti agricoli, le malattie, le vicende socio-politiche, il complesso dei rapporti Uomo-ambiente, etc. In tali ambiti, la mente umana registra regolarità senza certezze, convinzioni non sicurissime, ripetizioni di eventi non sempre garantiti da un esito univoco. Da un lato ciò genera paura e impone cautele contro i rischi (la mutualità prima e le assicurazioni poi), dall'altro invita al gioco e alla scommessa (inventando artificialmente l'aleatorietà nel risultato tramite semplici strumenti: palline, dadi, carte).

Pur essendo ben presente nella storia e nella cultura sin dai primordi della civiltà, la probabilità diventa un concetto importante e ben formalizzato solo a partire dal secolo XVIII quando sorge una nuova disciplina che sarà poi denominata: *Calcolo delle probabilità*. Tuttavia, nonostante importanti risultati formali, occorre attendere l'inizio del 1900

perché diventi palese e si diffonda lo studio dell'incertezza come fattore immanente nella scienza, nella tecnica, nelle attività umane. Questo avviene anzitutto nelle scienze biologiche, fisiche e mediche ma si trasferisce presto alle scienze umane, sociali, economiche e politiche, permeando – a partire dagli anni 1930/50 – tutta la ricerca scientifica, l'attività tecnologica e l'intero dominio della conoscenza razionale.

La saldatura tra queste due anime della Statistica è avvenuta solo quando, di fronte alla natura sempre più sperimentale della conoscenza, ci si è posti il problema della validità delle ipotesi.

A domande del tipo: «*Quanto è casuale il risultato cui si è pervenuti? quanto si può scommettere sulla validità di A rispetto a B?*» si può rispondere solo all'interno di una logica probabilistica. Così, il metodo statistico diviene nei fatti la metodologia della ricerca scientifica, essendo costantemente presente sia nelle analisi di laboratorio che nella ricerca di leggi generali: in sostanza, la Statistica diventa prassi scientifica ancor prima di essere riconosciuta come strumento autonomo di indagine.

Oggi, anche in conseguenza dei veloci mutamenti tecnologici ed informatici, si assiste ad un costante tentativo di utilizzare la Statistica (o le «statistiche») a sostegno di tesi predefinite, cioè come uno strumento di convincimento ideologico.

I sondaggi di opinione, importante tecnica statistica nata per esigenze conoscitive, possono divenire strumenti di manipolazione della realtà risolvendosi spesso in affermazioni senza riscontro metodologico e senza certificazione statistica. Difatti, al di fuori di un quadro di riferimento (e quindi di un modello conoscitivo), il sondaggio come messaggio punta a proporre l'esistente come legge, confondendo quindi ipotesi e tesi ma realizzando, nei fatti, il convincimento diffuso di una razionalizzazione empirica che trova in se stessa la norma e l'etica. Esiste quindi il rischio (che la tecnologia aiuta a nascondere) di un ritorno alla concezione della Statistica come emanazione del potere (ieri militare-statale, oggi tecnologico-informatico o semplicemente mediatico) e non come fattore di conoscenza umana e progresso scientifico.

È necessario, quindi, proporre una collocazione più equilibrata per la Statistica, nella quale siano visibili sia le radici storiche che i suoi contenuti scientifici, in modo che anche tale disciplina possa contribuire ad una migliore conoscenza della realtà e al miglioramento collettivo dell'Umanità.

1.3 La collocazione scientifica della Statistica

Talvolta, memori di una fase storica ormai superata, si cerca di collocare la Statistica nell'ambito dei metodi negando a tale disciplina il ruolo di scienza. In effetti, è complicato distinguere il momento creativo del metodo statistico dall'applicazione di strumenti consolidati (formali, computazionali, grafici) che di volta in volta vengono utilizzati nelle discipline positive (come, per esempio, nella Fisica o nella Biologia, nella Sociologia o nell'Economia).

D'altra parte, i metodi statistici presuppongono (in analogia a quanto avviene con la Logica e la Matematica) un rigore di analisi ed una logica ben formalizzata

che conferisce alla Statistica la natura propria delle scienze deduttive.

Per questo, ci pare corretto esaltare sia l'aspetto metodologico della Statistica che la natura innovativa delle proposte che tale disciplina mette in campo per migliorare l'approccio conoscitivo al mondo reale.

Tale conoscenza si concretizza in fasi e momenti successivi: definizione degli obiettivi, raccolta delle informazioni, elaborazione dei dati, presentazione dei risultati e decisioni operative conseguenti. Tuttavia, ogni corretta analisi statistica non è la giustapposizione sequenziale di tappe univoche (che, allora, potrebbero essere affidate ad una macchina) ma richiama, come è tipico della ricerca scientifica, uno schema teorico (concettuale ed esplicativo, quindi) che genera le osservazioni sperimentali.

1.4 Alcuni paradigmi della ricerca statistica

La Statistica utilizza, in modo esplicito o implicito, alcuni criteri logico-concettuali di riferimento (=*paradigmi*) nell'applicare la sua metodologia ai problemi reali. Alcuni di essi – prescelti in funzione del percorso scientifico che anima questo volume – verranno qui brevemente discussi.

- Il paradigma statistico di base è quello della *sintesi delle informazioni*. Di fronte all'analisi di fenomeni complessi – concernenti una pluralità di informazioni – si cerca un indicatore riassuntivo utile per qualche scopo predefinito e si attribuisce a quest'unica misura il significato di «rappresentanza».

Pur essendo evidente il rischio di tale «semplificazione», il pervenire ad un solo indice di fronte a molte informazioni costituisce una necessità concreta e quotidiana per assumere decisioni: per questo, la gran parte degli strumenti statistici presenti nella letteratura costituiscono proposte di sintesi che rispondono ad alcuni scopi e, quindi, proprio per questo, ne trascurano altri.

- Un secondo criterio statistico è la *scoperta del nuovo*, intendendo con ciò la messa in evidenza di fatti, congetture, problematiche, connessioni, leggi che non erano note in precedenza, o che andavano verificate. Tale paradigma si coniuga con l'atteggiamento sperimentale della ricerca scientifica moderna perché la metodologia statistica cerca di pervenire a risposte che hanno il pregio di essere associate ad una valutazione probabilistica di falsità, e non di certezza matematica. A tale riguardo, va sottolineato come lo statistico proceda per eccezioni, cioè ritenga più corretto rifiutare una ipotesi piuttosto che accettarne un'altra.

Questa impostazione critica e dialettica della conoscenza scientifica, tramite un'analisi soggetta ad errori, e quindi intrinsecamente statistica, deriva dal filosofo Karl Popper e si ritroverà con molto dettaglio nella parte dedicata alla logica del test delle ipotesi, assumendo così un ruolo centrale per tutta l'inferenza statistica. In questa scia, ritroviamo alcune affermazioni importanti: in particolare, che tutti i modelli sono sbagliati, che ogni apprendimento consiste nella eliminazione di un errore più che nella accettazione di una verità, che le soluzioni ai problemi quotidiani sono approssimazioni successive ottenute sulla base di esperimenti sequenziali, e così via.

- Un ulteriore paradigma è il *contenimento dell'errore*. In effetti, l'analisi statistica presuppone una ipotetica «scatola nera» (=*black-box*) nella quale esiste il vero legame tra i fatti ma che lo statistico non conosce né mai conoscerà, e che intende approssimare e accogliere dubitativamente sino a che una nuova proposta si presenti come più pregnante della precedente. La connessione tra le informazioni possedute in ingresso (=*input*) e quelle osservate in uscita (=*output*) è alterata da una componente erratica, di natura aleatoria e quindi imprevedibile.

Obiettivo di ogni ricerca scientifica è la riduzione della componente erratica (la sua eliminazione è solo una utopia o una falsificazione): più tale componente avrà un ruolo contenuto più la capacità di dominare la realtà ed utilizzare la conoscenza per fini prefissati diventa efficace.

- Infine, il paradigma essenziale della ricerca statistica è *la dialettica della conoscenza*, cioè l'accettazione dei risultati, la loro lettura critica, la proposta di ipotesi alternative e lo studio di nuovi risultati per proporre, rifiutare e migliorare precedenti acquisizioni.

L'aspetto iterativo ed interattivo nell'analisi dei dati con metodi statistici costituisce il cuore della ricerca statistica applicata, e ruota attorno alla capacità di costruire modelli che diano ragione sufficiente delle osservazioni e forniscano agli scienziati (come agli operatori/decisori) la concreta possibilità di migliorare (provvisoriamente) le loro teorie.

1.5 Problemi reali e problemi statistici

Nell'ambito della sua attività quotidiana, l'essere umano raccoglie informazioni sia per curiosità derivanti dalla sua sete di conoscenza che per raggiungere obiettivi concreti. In entrambi i casi, vale un principio generale che adotteremo per l'intero testo: *l'obiettivo specifica sia la natura delle informazioni da raccogliere che gli strumenti mediante i quali esaminare tali informazioni.*

Per questo, ogni analisi empirica – in modo esplicito o implicito – si articola secondo lo schema seguente:

Se, per esempio, obiettivo del decisore (Dirigente scolastico, Sindaco, Prefetto) è la predisposizione di servizi sociali sul territorio (scuola, sanità, sicurezza), una corretta indagine dovrà prioritariamente raccogliere nel modo più completo e fedele, tutte le informazioni utili connesse a quel particolare servizio (numero di famiglie con figli in età scolare, numero di anziani e frequenza di malattie per fasce di età, reati denunciati per

mese e tipologia). Questo momento è particolarmente delicato perché, per esempio, occorrerà definire se l'analisi riguarda i soli residenti o anche i presenti (ed in località turistiche ciò cambia notevolmente in taluni periodi dell'anno); se le malattie da studiare riguardano quelle infettive, di età pre-scolare, degli adulti, del periodo estivo (in funzione di tali scelte cambiano le modalità e i ritmi di rilevazione); se i reati sono solo quelli denunciati o quelli effettivi (è noto che taluni reati non vengono più denunciati perché ciò viene ritenuto inutile).

Successivamente, grazie ai metodi propri dell'analisi statistica (di cui ci occuperemo diffusamente nelle pagine seguenti), il decisore otterrà dei risultati che aiuteranno a specificare e rendere efficace il perseguimento degli obiettivi predefiniti, concretizzandosi in azioni ben visibili (reperimento di nuove aule, incremento del personale medico nelle unità sanitarie, adeguamento delle forze di polizia).

Infine, la valutazione dei risultati raggiunti ed i problemi evidenziati richiederanno, molto spesso, la ri-definizione di nuovi obiettivi, la raccolta di nuove informazioni, etc.

Come si vede, la metodologia statistica opera mediante stadi successivi (in ciascuno dei quali vi è iterazione del metodo ed interazione con i risultati) che implicano un atteggiamento teso al miglioramento progressivo della conoscenza. Questo risultato è più agevole se si evidenziano alcuni punti qualificanti:

- predefinire *obiettivi* delimitati ed agevoli da investigare sul piano empirico, in modo da consentire riscontri oggettivi, controlli accurati, verifiche regolari.
- individuare le *informazioni* necessarie per una puntuale conoscenza del fenomeno oggetto di studio, nonché predisporre le modalità di raccolta dei dati in modo che essa risulti completa rispetto ai soggetti interessati, efficace sul piano delle risposte, accurata nella codifica, efficiente rispetto ai costi e ai tempi dell'indagine;
- applicare *metodi* adeguati per il trattamento delle informazioni che consentano anche perfezionamenti mediante indagini preliminari;
- esaminare criticamente i *risultati* della metodologia statistica per saggiarne la coerenza, la validità, la generalizzabilità e, soprattutto, per verificare la loro utilità in rapporto alle domande poste dagli obiettivi dell'indagine.

Esaminiamo, ora, alcuni problemi reali che, pur essendo tra loro molto diversi, possono essere affrontati e risolti mediante l'applicazione di metodi statistici.

- Prevedere le piene di un fiume, migliorare la resa per ettaro di una coltivazione agraria, ottimizzare i tempi di attesa ad uno sportello bancario o ad un casello autostradale, produrre medicinali efficaci con minori effetti collaterali dannosi, investire risparmi in azioni meno oscillanti delle altre nel mercato borsistico, migliorare l'efficacia di metodi didattici, valutare gli effetti della televisione sul comportamento degli adolescenti o l'impatto dei messaggi pubblicitari sulle decisioni di acquisto dei consumatori, etc. etc. sono solo alcuni degli esempi delle situazioni reali nelle quali la Statistica interviene per agevolare l'assunzione di decisioni corrette e coerenti. Peraltro, ogni disciplina particolare si struttura con accentuazioni proprie per utilizzare la Statistica in modo efficiente.

- Lo studio delle piene dei fiumi si configura come un monitoraggio costante nel quale la previsione in tempi ravvicinati è elemento decisivo nella scelta tra strumenti alternativi di intervento e/o allarme.
- La sperimentazione agraria e quella clinica possono utilizzare metodi statistici assai simili ma nelle indagini sulla salute umana diventano prioritari i fattori etici, la minimizzazione dei rischi, le soglie di sopportabilità del dolore: tali aspetti sono decisivi nella scelta delle metodologie.
- Gli investimenti azionari seguono ritmi giornalieri, e persino orari, mentre gli affollamenti ai caselli autostradali e le code agli sportelli bancari o postali sono molto più sensibili alle scadenze, ai giorni della settimana, alla loro collocazione sul territorio.
- I messaggi televisivi e quelli pubblicitari inducono comunque dei comportamenti sia sugli adolescenti che sui consumatori, ma per queste ultime tipologie il marketing ha raffinato strategie e tecniche molto più articolate che per il generico comportamento dei più giovani: infatti, la pubblicità mira ad effetti specifici, delimitati nel tempo e controllabili tramite i consumi, mentre gli effetti complessivi della televisione su un pubblico giovanile sono meno specifici, variabili nel tempo e difficilmente isolabili e misurabili.

Questi esempi mostrano che le indagini statistiche perseguono obiettivi molto diversi, ma la natura del fenomeno delimita condizioni al contorno, elementi aggiuntivi, modalità di acquisizione dei dati, vincoli ineliminabili, tempi di utilizzo dei risultati, e così via. D'altra parte, sono spesso i problemi reali a suggerire lo sviluppo di metodi statistici innovativi i quali, a loro volta, propongono nuove soluzioni per quegli stessi problemi.

Da un certo punto di vista, la fase più delicata della ricerca statistica e delle applicazioni della Statistica ai problemi concreti consiste proprio nella corretta trasposizione di un problema reale (quello che sta a cuore al ricercatore) in un problema statistico (quello la cui soluzione è ben codificata nei testi).

Il passaggio $\boxed{\textit{problema reale} \rightarrow \textit{problema statistico}}$ richiede esperienza specifica e cultura statistica e può avvenire in modo agevole solo mediante la interazione tra studiosi di discipline differenti.

1.6 Fasi di una analisi statistica

È ora opportuno esaminare in dettaglio le fasi di una indagine statistica; infatti, esse richiedono una accurata pianificazione allo scopo di consentire alle informazioni raccolte di essere analizzate in modo appropriato con metodi statistici.

1. ***Definizione degli obiettivi della ricerca.*** Gli obiettivi devono essere molto particolareggiati per individuare con cura le informazioni da ricercare, evitando equivoci nelle definizioni, circoscrivendo con esattezza il territorio e il periodo della indagine, fornendo soluzioni alternative nei casi di mancate o errate risposte, corresponsabilizzando rilevatori o intervistatori per migliorare la qualità dei dati raccolti. Spesso, occorrono studi preliminari per circoscrivere l'ambito di una indagine e tutte le possibili modalità di risposta.

2. ***Rilevazione dei dati.*** La rilevazione dei dati può essere *completa* (quando si esaminano tutti gli elementi oggetto di studio) o *parziale* (quando ci si limita a studiare un sottoinsieme, detto «campione», dell'insieme di riferimento): questi aspetti saranno sviluppati più oltre, nel testo. A tale proposito, esistono seri problemi circa la numerosità degli elementi da studiare, la lista delle informazioni statistiche (o dati) da raccogliere, la tipologia delle risposte o delle misurazioni ammissibili, gli strumenti di rilevazione, le persone cui demandare tale compito, i costi della rilevazione. Tra questi aspetti, è essenziale esplicitare il modo con cui si raccolgono le informazioni (e cioè, tramite *dichiarazioni* – mediante questionario, intervista, registrazione, ovvero tramite *misurazioni* – mediante strumenti semplici o complessi) perché le modalità di raccolta dei dati determinano l'accuratezza delle informazioni su cui successivamente si dovrà operare.

3. ***Elaborazione metodologica.*** In questa fase si applicano schemi formali di tipo logico-deduttivo e inferenziale-induttivo, di natura matematica, cioè gli strumenti propri dell'analisi statistica, alcuni dei quali saranno presentati, discussi ed applicati nei Capitoli successivi. Anche se è difficile definire una volta per tutte le metodologie più idonee (anche perché una pluralità di metodi sugli stessi dati consente di esaminarli da differenti punti di vista), la distinzione tra fenomeni «qualitativi» e «quantitativi» (sulla quale ci soffermeremo più oltre) gioca un ruolo decisivo per la scelta dei metodi che si possono applicare. Inoltre, di fronte a problemi nuovi e complessi, lo statistico può anche sviluppare metodologie completamente innovative.

4. ***Presentazione ed interpretazione dei risultati.*** Una presentazione poco efficace può vanificare sia l'utilità dei risultati raggiunti che le potenzialità del metodo utilizzato. Invece, una accurata illustrazione dei risultati (sotto forma di tabelle, grafici, diagrammi di relazioni, rapporti sintetici) e una disamina particolareggiata delle implicazioni operative dei medesimi, con particolare riferimento alle interpretazioni proprie del settore in cui si opera, sono elementi decisivi per il buon esito di una indagine statistica.

5. ***Utilizzazione dei risultati della ricerca.*** Lo statistico contribuisce ad una corretta utilizzazione operativa dei risultati di una indagine scientifica, circoscrivendo l'ambito interpretativo e richiamandosi ai vincoli entro cui essa assume validità. L'uso che può essere fatto dei risultati di una ricerca deriva, quindi, anche dalle sue modalità di esecuzione e dai dettagli mediante i quali è stata condotta la misurazione o l'intervista (occasioni, tempi, luoghi, etc.). Infatti, solo precisando questi aspetti è lecito ampliare l'interpretazione dei risultati.

Infine, pur pianificando con cura ogni aspetto di una ricerca, la realtà fenomenica è così complessa ed articolata che, regolarmente, ci si trova di fronte a situazioni non previste e alla necessità di modificare numerosi aspetti dell'indagine. Se ciò può creare problemi nella gestione e previsione di tempi e risorse, esso è anche un connotato positivo della ricerca empirica. Infatti, interagendo con la realtà in modo intelligente si scoprono fatti insospettabili e nuove relazioni tra fenomeni, il che produce un sicuro arricchimento della conoscenza.

1.7 Il percorso logico di questo volume

Questo volume si snoda secondo un percorso logico ed unitario che conferma il ruolo della Statistica come scienza. Tale percorso trae origine dalle analisi esplorative dei dati disponibili (parte prima), trova una formalizzazione razionale mediante gli strumenti probabilistici (parte seconda) ed inferenziali (parte terza) per pervenire, infine, alla costruzione di modelli utili ed efficaci per le decisioni (parte quarta).

Tale sviluppo fa riferimento al processo conoscitivo dell'Uomo nel suo rapportarsi alla realtà esterna e si concretizza in tappe di apprendimento, funzionali all'obiettivo che intende raggiungere. Nella sua essenza, la Statistica interviene in ogni fase del pensiero e dell'azione e si articola secondo la sequenza deduttiva che dalla percezione sensoriale passa alla riflessione cosciente e si concretizza nell'azione mediante strumenti.

- La prima percezione dell'esistente avviene mediante il senso della vista intesa come cattura della realtà (nella varietà e mutevolezza, per analogia e similarità), allo scopo di confrontare e sintetizzare fenomeni, eventi, problemi nella loro individualità e/o nella loro complessità. In tal senso, la Statistica è scienza che aiuta a *vedere* il mondo, a fotografarlo, a catalogarlo e classificarlo cogliendo, misurando ed esplicitando l'essenziale, il caratterizzante, ciò che fa la differenza.

- L'atto successivo al vedere è la riflessione e la comprensione, cioè *capire* la realtà all'interno di una impostazione probabilistica nella quale l'esistente è esaminato in rapporto a ciò che poteva accadere o che verosimilmente accadrà. L'indeterminismo come aspetto ontologico del reale e la probabilità come sua misura costituiscono l'atteggiamento e il sapere dello scienziato moderno. Infatti, una conoscenza al di fuori della logica del probabile, nelle scienze sociali come in quelle fisiche e naturali, è una mediocre approssimazione alla realtà e rappresenta un residuo di atteggiamenti dogmatici.

  Peraltro, la comprensione di una fenomenologia soggetta ad incertezza presuppone sia il *momento deduttivo* (chiedersi cosa avverrà sulla base di schemi noti) che il *momento induttivo* (chiedersi quale meccanismo incognito può aver provocato i risultati empirici). Questi due momenti fanno riferimento all'aspetto deduttivo (probabilità) ed induttivo (inferenza) insito in qualsiasi riflessione e sono entrambi dettati dalla ineliminabile componente di incertezza che circonda qualsiasi conoscenza umana.

- Infine, senza l'azione l'Uomo non sopravvive né migliora. Esistono obiettivi da perseguire, progetti da realizzare, «cose da fare», la cui operatività deriva dalla comprensione, ma implica anche una progettualità: questa fase è detta dell'*agire* per raggiungere scopi predefiniti e consentire la fruibilità del mondo, della società, delle risorse e delle opportunità che si offrono.

La descrizione (il vedere) e la comprensione (il capire) orientate verso l'azione concreta e l'obiettivo specifico (l'agire) producono sempre un modello statistico che svolge un ruolo strumentale, circoscritto, limitato ma utilissimo ed ineliminabile. Per questo, anche quando non viene esplicitamente detto, la modellistica

dei fenomeni reali è il momento finale della ricerca statistica proiettata verso le applicazioni.

Pur essendo soggettivo, e quindi opinabile, questo percorso ci ha aiutato a guardare con curiosità e con passione alla Statistica come scienza logicamente coerente, sia nel suo impianto teorico che nelle sue numerose verifiche: soprattutto, questo percorso unifica molti aspetti della Statistica all'interno di un apprendimento unitario e ben finalizzato.

1.8 Considerazioni finali

La Statistica come scienza e metodo della ricerca è strumento di conoscenza e di progresso anche se, come tutti gli strumenti prodotti dall'Uomo, può essere utilizzata in modo incorretto o improprio, precostituendo i risultati di una indagine o orientandone le interpretazioni.

Così, come in tutte le scienze, anche nella Statistica si possono ritrovare risultati scientifici notevoli dedotti con metodi rigorosi, ma anche autentiche falsificazioni derivanti da faciloneria, ignoranza o malafede. In tempi recenti, è divenuto frequente l'uso della *Statistica come spettacolo* mediante la quale, soprattutto nei mass-media, si cerca con dati raccolti in modo non-scientifico di forzare una proposta di vendita, un comportamento sociale o un orientamento politico.

In effetti, i dati e le informazioni non costituiscono verità assoluta ma solo realizzazione di una legge o di una ipotesi che quei dati devono mettere in discussione. *Senza modelli falsificabili dall'esperienza, i dati statistici sono numeri privi di senso.* Al di fuori di uno schema concettuale rigoroso, non esiste interesse scientifico né utilità pratica per qualunque analisi statistica. Per questi motivi, comunicare la percentuale dei favorevoli di fronte ad una affermazione A non costituisce una informazione statistica, corretta ed utilizzabile, sino a quando essa non viene formalizzata e interpretata nell'ambito di un esperimento ben definito e soggetto ad incertezza.

Per tutto questo, a noi sembra corretto definire la Statistica come la *Scienza delle decisioni in condizioni di incertezza.*

Ambiente R
Statistica e metodologia della ricerca

R1.1 Cosa è l'ambiente statistico R

In queste pagine verranno mostrate le principali istruzioni necessarie per chi intende esercitarsi sui data set che accompagnano il libro ovvero su propri dati, replicando le analisi statistiche o sviluppandone altre.

A tal fine, *queste note non intendono essere una introduzione alle molteplici potenzialità dell'ambiente* R perché esse sono funzionali alle possibilità applicative di un corso base di Statistica. Per approfondire questi aspetti ci si può riferire ai testi che approfondiscono l'argomento (un elenco è nel paragrafo 1.7), nonché al vasto materiale disponibile gratuitamente in Internet. Peraltro, le potenzialità di tale ambiente sono di gran lunga più estese di quello che potrebbe apparire da questi appunti. Soprattutto, queste note introducono progressivamente (e non in modo sistematico) le potenzialità di R in funzione delle esigenze di calcolo e di rappresentazione grafica per un approfondimento della Statistica secondo il percorso delineato nel testo.

Storicamente, pur essendosi evoluto sulla base di software e linguaggi precedenti (in particolare, il linguaggio *S*, poi diffuso in commercio come *S-Plus*, e di cui John Chambers va citato come un precursore), R è stato proposto da Ross Iahaka e Robert Gentleman del Department of Statistics, University of Auckland, New Zealand, a partire dal 1996, in collaborazione con altri statistici, ed è attualmente aggiornato e curato da uno staff internazionale (*R Core Development Team*).

In breve, R è sia un software che un linguaggio di programmazione (talvolta, si definisce *un ambiente statistico*). Rispetto ai tanti software e linguaggi di programmazione esistenti (liberi e commerciali), la rapida e tuttora crescente diffusione di R fra gli statistici è dovuta ad alcuni fattori:

- R è un software disponibile gratuitamente, immediatamente utilizzabile su quasi tutti i sistemi operativi più diffusi, e dotato di potenti strumenti di calcolo che consentono semplici e complesse analisi statistiche nonché malleabili e sofisticate rappresentazioni grafiche. Ciascuno può utilizzarlo, modificarlo e redistribuirlo, purché con la stessa licenza (GNU) con la quale lo ha ricevuto.

- R è accompagnato da istruzioni standard e di facile uso ma consente anche una programmazione avanzata per utenti più esperti. In ogni caso, permette una personalizzazione molto marcata delle analisi e dei risultati.

- R si presenta con una struttura di base che si amplia continuamente grazie alla disponibilità di programmi specifici (i *packages*, la cui numerosità raggiunge già il migliaio) che sono messi a disposizione gratuita dalla comunità internazionale di studiosi e ricercatori. Gli statistici, oggi, per la gran parte, comunicano e si confrontano tramite questo ambiente.

Molti software statistici consentono analisi predefinite (talora, anche sofisticate) ma non sempre permettono all'utente di modificare input ed output in funzione delle proprie esigenze. Invece, R è completamente programmabile in funzione degli obiettivi della ricerca. Ciò implica che, di per sé, R richiede all'utente di immettere di volta in volta le istruzioni per gli obiettivi che vuole conseguire. Quindi, all'inizio dell'esperienza *non si apprezza appieno* la potenza di analisi statistica e grafica di R ; tuttavia, non appena si entra nella logica di tale ambiente è difficile poi ritornare ad utilizzare i software tradizionali. Ciò è dovuto al fatto che, all'inizio, l'utente ha l'impressione di dover immettere le istruzioni una alla volta, anche per ottenere risultati banali (una semplice media, un grafico elementare); poi, apprende che le istruzioni possono essere raccolte e riutilizzate con semplici varianti, ed adattate alle specifiche nuove esigenze (se necessario). Già dopo i primi Capitoli di queste note, questa abilità di R verrà apprezzata dal lettore.

Per scaricare gratuitamente l'ultima versione aggiornata di R , basta collegarsi al sito: `http://cran.r-project.org/`

Gli aggiornamenti delle versioni avvengono di norma con ritmo semestrale. Attualmente, R è disponibile per numerosi sistemi operativi, ma queste note si riferiscono a chi possiede un PC con sistema operativo Windows (essendo questo il più diffuso tra i potenziali lettori). Quindi, occorre scaricare la versione di R eseguibile nel sistema operativo Windows.

R1.2 Come si attiva una sessione di R

Supponendo sia stato installato correttamente (la procedura è molto semplice e basta seguire le istruzioni, disponibili anche in lingua Italiana), il programma si attiva facendo clic sulla icona di R presente sul desktop. In tal modo, si apre una finestra entro la quale digitare tutte le istruzioni successive: tale finestra di R (che l'utente può personalizzare in dimensioni, fonti, colore) viene utilizzata sia per immettere istruzioni che per ricevere risultati (a meno che non sia codificato altrimenti oppure che il risultato non sia un grafico: in tal caso si apre anche una finestra grafica). In particolare, in alto sono accessibili (mediante l'uso del mouse) numerosi comandi per gestire file, directory, informazioni, pacchetti e richiedere aiuto.

Il prompt di R è > e può essere modificato dall'utente: in questo libro, è stato preferito `R==>` per migliorare la visibilità delle istruzioni. Ogni istruzione viene attivata utilizzando (dopo averla digitata) il tasto `Invio`. Se più istruzioni (*statements*) sono sulla stessa linea, occorre separarle tramite ; In queste note si userà (quasi sempre) concludere le istruzioni con ; anche se ciò non è strettamente necessario. Questa scelta, a nostro parere, migliora la comprensione e la leggibilità delle singole istruzioni. Quindi, d'ora innanzi, ogni istruzione che l'utente deve immettere da tastiera è preceduta da `R==>` mentre l'output del sistema è quello che segue.

Entrando in R , la sessione di lavoro si svolge in una *working directory* predefinita dall'ambiente (può essere configurata dall'utente di volta in volta oppure mediante i file che definiscono il profilo iniziale). Quando si desidera che la sessione si svolga in una specifica *directory*, per esempio `librostat`, occorre digitare (all'inizio della sessione di lavoro e prima di ogni altra istruzione):

```
R==> setwd(librostat)
```


Tale risultato può anche essere ottenuto andando con il mouse nel menù in alto a sinistra (nella finestra iniziale di R) e scorrendo sotto "File" sino a trovare "Cambia directory ..." e poi scegliere quella desiderata.

È preferibile che l'intera attività (di studio, di ricerca, di esercitazione) si svolga entro una specifica *directory* sia per conservare i dati utili che per risparmiare la digitazione dei nomi dei files da richiamare; se, infatti, i file sono in una *directory* differente rispetto a quella di lavoro occorre digitare il nome completo. Per esempio, un file che si chiama `miofile.text` sarà richiamato mediante: `''miofile.text''` quando si è già nella sua *directory*; altrimenti, esso va richiamato mediante la sua designazione completa. Per esempio, con:

`''C:\\Program Files\\R\\R-2.6.2\\librostat\\miofile.txt''`

Si osservi come l'ambiente R richieda nella definizione del percorso (*path*) di un file l'uso dei simboli \\ (oppure /) per definire correttamente le *directory*.

Nel dubbio, l'istruzione `getwd()` fornisce la *directory* in cui si sta lavorando.

Tutto quello che segue su una linea dopo il simbolo # viene inteso come "commento" e, quindi, non viene eseguito da R . Questa caratteristica viene utilizzata per commentare istruzioni ed output (come noi faremo, spesso).

Tramite la linea di immissione, il sistema conserva le istruzioni attivate per l'esecuzione; quindi, le istruzioni precedenti si possono richiamare mediante i tasti delle frecce: ↑ e ↓. Invece, le frecce ← e → sono utili per ri-editare la linea in R .

Se si vuole conservare in un file ciò che il sistema invia a schermo basta porre all'inizio e alla fine della sessione la funzione `sink()` nel modo seguente:

```
R==> sink("miofile")      # da questo punto, tutte le elaborazioni
                          # vengono conservate nel file "miofile"
R==> ::::::::::::::       # lista di istruzioni di R
R==> sink()               # da questo punto, nulla è più conservato
```

Il file `miofile` sarà un file di testo, quindi utilizzabile per ambienti di scrittura e/o di elaborazione.

Entrando in R , alcuni programmi (*packages*) statistici e grafici di base vengono automaticamente caricati; quando sono necessari pacchetti più specifici occorre attivarli mediante l'istruzione:

```
R==> library(newpack)     # si richiama il package "newpack",
                          # se è già disponibile nelle librerie
                          # caricate da R nel sistema
```

In alternativa, si può anche utilizzare l'istruzione: `require(newpack)`. Se il *package* in questione non è già disponibile sul proprio PC, occorre installarlo (da uno dei siti disponibili) utilizzando i comandi in alto sulla finestra di R (occorre connettersi ad Internet per scaricare tali pacchetti). Se l'utente ritiene di dover caricare taluni pacchetti ogni volta che entra in R (in modo automatico) è sufficiente modificare il file iniziale che designa il profilo di R : queste modalità sono descritte nei manuali e non sono necessarie per il seguito.

Per la lista di tutti i *packages* disponibili occorre consultare il sito di R , che contiene anche una loro sintetica descrizione. A tal fine, giova ricordare le seguenti funzioni:

```
R==> list.files()             # elenca tutti i file della
                              # working directory
R==> installed.packages()     # elenca i packages di R
                              # installati sul PC
R==> available.packages()     # elenca i packages disponibili tramite
                              # collegamento ad uno dei siti di R
```

Se, nelle precedenti istruzioni, invece delle parentesi () si specifica una *directory*, le istruzioni si applicano a quella *directory*.

L'istruzione `ls()` elenca tutti gli oggetti presenti nella *directory* di lavoro mentre `rm()` ne elimina quelli specificati; si vedano le esemplificazioni per il loro uso congiunto:

```
R==> rm(a,b,d)        # cancella a, b, d dalla working directory
R==> rm(list=ls())    # cancella TUTTI gli oggetti presenti
                      # nella working directory
```

L'ultima istruzione è utile quando si intende ri-cominciare una sessione di lavoro senza correre il rischio di utilizzare (inavvertitamente) oggetti già presenti; tuttavia, va usata con cautela perché essa cancella *tutti* gli oggetti in modo definitivo.

Per uscire da R , si fa clic con il mouse sinistro sulla [×] posta in alto a destra della finestra, oppure si digita la funzione `q()`. Prima di uscire da R , il sistema chiede se si desidera salvare la sessione di lavoro, ed è usuale rispondere "no".

R1.3 Richiesta di aiuto in R

Per ricevere aiuto su di una funzione di cui si conosce il nome, si possono utilizzare le seguenti istruzioni (equivalenti):

```
R==> help(mean);help("mean"); ?mean # aiuto sulla funzione "mean"
```

Nel caso di operatori (come di seguito esemplificato) sono necessarie le "":

```
R==> help("["); ?"["        ### aiuto sull'uso delle parentesi [
R==> help("%*%"); ?"%*%"    ### aiuto sull'uso dell'operazione %*%
```

Si può chiedere aiuto anche sulle funzioni di qualsiasi *package* attivo in R .

Si può chiedere anche di trovare in quale pacchetto è utilizzata una certa espressione; in tal caso, si utilizza:

```
R==> help.search(histogram)       # aiuto sulla parola "histogram"
```

R1.4 Nozioni di base per lavorare in R

Da un certo punto di vista R esegue operazioni, calcola funzioni matematiche ed applica formule algebriche come un qualsiasi calcolatore numerico, utilizzando le regole dell'algebra ordinaria (mediante simboli e funzioni di chiaro significato, sulle quali non ci soffermiamo). Si osservi come la risposta sia preceduta da `[1]` quando trattasi di un solo valore.

```
R==> 3+4                        # addizione di due numeri interi
[1] 7

R==> 2-10; 3/4; 5*10+2*(9-12)^2-1;  # operazioni algebriche
[1] -8
[1] 0.75
[1] 67

R==> log(10);                   # logaritmo (neperiano) di 10
[1] 2.3026
R==> sqrt(12);                  # radice quadrata di 12
```



```
[1] 3.4641
R==> cos(2*pi)                  # coseno del numero 2*(pi-greco)
[1] 1
R==> factorial(6);              # fattoriale di 6 = 1*2*3*4*5*6
[1] 720
R==> choose(5,3);               # numero di combinazioni di
                                # 5 oggetti a 3 a 3
[1] 10
```

L'ultima istruzione calcola il coefficiente binomiale $\binom{5}{3}$; in generale, l'istruzione è:

`choose(n,k)`.

Ovviamente, R rispetta le regole algebriche di precedenza delle parentesi, come evidenziato nelle esemplificazioni seguenti:

```
R==> 2+4*(3-1);
[1] 10
R==> 2+4*3-1;
[1] 13
```

Si noti che R utilizza il simbolo (riservato) `pi` per la costante matematica $\pi = 3.141593....$ Esistono altri simboli riservati (l'utente viene avvisato per un uso improprio) e tra questi citiamo `c()` (che opera una concatenazione) e `t()` (che opera la trasposizione) nonché `F` (che sta per `FALSE`) e `T` (che sta per `TRUE`), che sono valori logici di cui discuteremo più avanti.

Si ricordi che R distingue tra lettere maiuscole e minuscole (si dice che è *case sensitive*); quindi, le espressioni `Questo; questo; QUESTO; quesTO; QUEsto;` fanno riferimento ad oggetti tutti differenti. Invece, quasi sempre, gli spazi tra istruzioni non creano problemi ad R .

In R l'assegnazione (*assignment*) del numero 3 alla variabile x, per esempio, si ottiene tramite l'istruzione `x <- 3` oppure tramite `x=3`. Invece, la verifica logica (che assegna "vero" `TRUE`, oppure "falso" `FALSE` alla variabile y) avviene tramite `y <- (4-1==3)` (che assegna ad y il valore `TRUE`) ovvero `y <- (4-1=5)` (che assegna ad y il valore `FALSE`).

"Assegnazione" significa attribuire il valore alla destra del simbolo `<-` alla variabile posta alla sinistra del simbolo `<-` . Quindi, se si è stabilito in precedenza che: `x <- 3`, l'assegnazione `x <- x+1` è corretta e serve ad attribuire, poi, il valore 4 alla variabile x. Viceversa, se si interpreta l'assegnazione come una banale uguaglianza aritmetica, allora l'istruzione `x <- x+1` non ha alcun senso, essendo impossibile che x sia uguale a $x + 1$, per ogni x finito.

Ciò premesso, poiché nel seguito non vi è rischio di confusione, d'ora innanzi tutte le assegnazioni di R saranno in questo testo *sempre* indicate con il simbolo = (che è formalmente corretto in R , ma viene sconsigliato dalla maggior parte degli utilizzatori) perché didatticamente questo semplifica la lettura delle istruzioni. Questo criterio rende le istruzioni più immediate ed è anche utilizzato in alcuni testi diffusi a livello internazionale, come Verzani (2005), Spector (2008) e Gentleman (2009).

Si osservi che l'assegnazione è una istruzione che pone il risultato di una elaborazione in una variabile. Se si desidera conoscere il risultato dell'assegnazione occorre richiederla, come nell'esempio seguente:

```
R==> z=3+4+7; z     # poni in z il calcolo (3+4+7), e poi mostra z
[1] 14
```

Per facilitare la comprensione di questo aspetto, basti pensare che il risultato di un'istruzione o viene posto in una variabile oppure, in assenza, viene inviato allo schermo. Si osservi la differenza tra le due seguenti istruzioni:

```
R==> z=3+4+7; z    # poni (3+4+7) in z, e poi mostra z
[1] 14
R==> 3+4+7         # esegui (3+4+7), ed invialo allo schermo
[1] 14
```

Nella prima istruzione, il risultato 14 è conservato in z (fino a che non si termina di lavorare in R ovvero si definisce z in modo diverso); nella seconda istruzione il medesimo risultato è inviato allo schermo (ed è quindi perso, perché non è memorizzato in nessuna variabile). Se, poi, si vuole ottenere sia l'assegnazione che il risultato a schermo basta porre l'istruzione entro parentesi, come nell'esempio:

```
R==> (z=3+4+7);       # poni (3+4+7) in z, e poi mostra z
[1] 14
```

Nei casi in cui si desidera esplicitamente visualizzare un risultato nella finestra di R (cioè a video) si utilizza la funzione `print()`:

```
R==> print(z);
[1] 14
```

Il numero di decimali visualizzato in R (non quello effettivamente utilizzato nei calcoli) è stabilito dall'ambiente e può essere modificato una volta per tutte nel profilo iniziale, all'inizio della sessione di lavoro, oppure quando si cerca di visualizzare una variabile numerica con una prefissata approssimazione, come esemplificato nelle istruzioni seguenti, mediante la funzione `round()`:

```
R==> pi; round(pi, digits=3);round(pi, digits=5);
[1] 3.1416
[1] 3.142
[1] 3.1416
```

Come si vede, la gestione interna di tale funzione in R non è sempre coerente con le richieste dell'utente. Tuttavia, per chi volesse "scoprire" le ulteriori cifre decimali che R utilizza, per esempio, per il calcolo di π si può usare il seguente "trucco":

```
R==> pi-3.1415;
[1] 9.2654e-05  ### le cifre dopo 3.1415 sono "92654"
R==> pi-3.141592653;
[1] 5.8979e-10  ### le cifre dopo 3.141592653 sono "58979"
```

e così via. Quindi, per l'ambiente R , $\pi \simeq 3.14159265358979...$

La funzione `View(cosa)` consente di visualizzare il contenuto di cosa se questa è assimilabile ad un oggetto contenente righe e colonne (come un foglio di calcolo) oppure un elenco. Invece, l'istruzione: `cosanuova=edit(cosa)` consente di modificare e salvare il contenuto dell'oggetto cosa in nuovacosa, attivando un editore molto semplice (trattasi di *Notepad* se non specificato diversamente). Si osservi che, in tal modo, in memoria rimane sia l'oggetto precedente che quello modificato, a meno di non ricoprirlo con una istruzione del tipo: `cosa=edit(cosa)`. Finalmente, l'istruzione `fix(cosa)` consente sia di editare che di modificare e conservare il contenuto di cosa.

R1.5 Lavorare in R

Ciascuno può eseguire una dopo l'altra le istruzioni in R e ricordarle a mente o scriverle da parte, se deve ripeterle. Se però occorre rifare le stesse (o simili) analisi, e si vuole evitare di digitare nuovamente quelle istruzioni, è utile conservarle in un file, chiamato *script* (=sceneggiatura, copione, documento).

Per creare e salvare in un file ''mioscript'' le istruzioni che si è interessati a conservare direttamente in R (l'editore si attiva facendo clic in alto a sinistra sulla finestra di R e scegliendo dal menù prima "File" e poi "Nuovo script" oppure "Apri script", se già esiste), si scrive nel file quanto desiderato, salvandolo e poi eseguendolo, quando necessario (con o senza modifiche) mediante la funzione `source("mioscript")`.

Anche l'editore Notepad è adeguato allo scopo per comporre il file *script* mediante istruzioni, ma ne esistono di specifici per R (per esempio, Tinn è consigliabile tra quelli gratuiti). Per produrre queste note, gli *script* sono stati costruiti ed utilizzati mediante il *package* R-WinEdt, che si scarica gratuitamente da R (ma richiede a sua volta il *software* WinEdt, molto efficiente ed essenziale per chi già utilizza LaTeX, acquistabile con un costo simbolico). In ogni caso, *non utilizzare mai Word.*

Essendo tale aspetto essenziale per ottimizzare (e semplificare di molto) l'uso di R da parte di qualsiasi utente, schematizziamo la procedura da seguire per chi volesse creare uno *script* per la prima volta:

1. Si apre un file (=*script*) che si intende poi sottoporre ad R : ciò avviene facendo clic in alto a sinistra, nella finestra di R , su "File" e poi su "Nuovo script".

2. Si scrivono, come testo, le istruzioni di R andando daccapo per ogni nuova istruzione (oppure, separandole con ;). R dispone di un editore essenziale per scrivere le istruzioni.

3. Si sottopongono ad R le istruzioni prescelte (o anche tutte) selezionandole e poi facendo clic con il tasto destro del mouse e scegliendo "Esegui linea o selezione". Si può anche evidenziare e poi applicare "Ctrl+R" (cioè premendo contemporaneamente sia il tasto "Ctrl" che "R").

4. Se si è soddisfatti dell'esito, le istruzioni si salvano selezionando con il mouse, nell'editore dello *script*, prima "File" e poi "Salva" (oppure "Salva con nome", designandolo per esempio come `mioscript`).

5. Per ogni utilizzo successivo di tale *script* (e delle istruzioni che contiene), basta aprirlo in R , oppure digitare l'istruzione `source("mioscript")` per eseguirlo tutto intero.

Occorre fare attenzione a richiamare tale file in modo corretto perché, in assenza di estensione, R lo designa come `mioscript.R`; in alternativa, può essere salvato come: `mioscript.txt`.

R1.6 I data set utilizzati per le analisi

Tutte le applicazioni che proporremo con R (e buona parte di quelle presenti nel testo) sono condotte su insiemi di dati reali (definiti *data set*) che presentano caratteristiche differenziate in modo da facilitare il lettore nella ricerca di problematiche affini alle sue. Tali dati sono scaricabili dal sito web del libro, all'indirizzo **www.mulino.it/aulaweb.**

Spesso, un'analisi che si discute non è importante di per sé ma diventa rilevante per analogia: per esempio, può sembrare poco interessante approfondire (mediante una corretta analisi statistica) la discriminazione tra le specie di *Camellia* mediante l'analisi di misure e peso delle foglie. Tuttavia, l'esempio è immediatamente trasferibile in ambito sanitario (per una corretta diagnosi), economico (per la misura dei rischi di investimento), ambientale (per la valutazione degli effetti di inquinanti), politico (per discutere di politiche dei redditi), e così via.

Inoltre, i dati sono connessi a situazioni che ciascuno può dimensionare senza sforzi ulteriori, replicare nel suo ambiente di lavoro, studio o tempo libero e sono tutti già disponibili in rete e predisposti in modo da rendere immediate le analisi in R .

Ogni data set è indicato con una sigla (che sarà richiamata fedelmente in queste note e nel testo) ed è qui di seguito brevemente descritto:

PREZZOLIBRI È una raccolta dei prezzi in Euro dei 420 libri esposti in un catalogo di un importante editore nazionale nel corso del 2001 (`prezzolibri.txt`).

CAMELLIA Contiene misure fisiche su alcune variabili relative a 333 foglie di *Camellia japonica* (la comune camelia da giardino). Per ciascuna foglia sono disponibili la lunghezza, la larghezza, il peso della foglia appena colta (=`pesovivo`) e dopo un essiccamento di circa 30 giorni (=`pesosecco`), il codice della specie (`camellia.txt`).

STUDENTI Sono i risultati ottenuti per gli anni 1988-1995 sulle risposte dei questionari compilati dagli studenti alla fine del corso di Statistica della Facoltà di Scienze Politiche dell'Università di Napoli Federico II. Contiene le variabili anno, genere, età (in anni), peso (in kg), altezza (in cm), diploma secondario di provenienza, numero di componenti della famiglia, occhiali (sì/no), fumo (sì/no) (`studenti.txt`).

COMPACT Contiene anno di edizione e durata di 1192 brani musicali presenti in una collezione di *compact disc* (`compact.txt`).

PRODUZIONE È la serie mensile dell'indice ISTAT della produzione industriale in Italia (corretta per tenere conto dei giorni lavorativi di ciascun mese), dal Gennaio 1990 a Dicembre 2003, consistente in 168 osservazioni consecutive (`produzione.txt`).

R1.7 Riferimenti bibliografici per R

In primo luogo, si consiglia di consultare il sito di R , esaminando il numeroso materiale disponibile sia in Italiano che Inglese.

Dopo aver installato R sul proprio PC, sono disponibili i file pdf contenente i manuali in Inglese (cliccando in alto su "Aiuto").

Per chi desidera testi più specifici occorre distinguere tra libri di Statistica che utilizzano R in modo sistematico e libri di R nei quali, ovviamente, esistono molte utilizzazioni statistiche (ma anche matematiche, numeriche, etc.).

A tal fine, riteniamo utile proporre una lista che, pur se non esaustiva, è quella di cui abbiamo conoscenza al momento di andare in stampa.

Albert, J. (2007) *Bayesian Computation with R* , New York, Springer.

Baclawski, K. (2008) *Introduction to Probability with R* , Boca Raton, FL, Chapman & Hall/CRC.

Braun, W.J. e Murdoch, D.J. (2007) *A first Course in Statistical Programming with R* , Cambridge, Cambridge University Press.

Chambers, J.M. (2008) *Software for Data Analysis. Programming with R* , New York, Springer.

Cohen, Y. e Cohen, J.Y. (2008) *Statistics and Data with R* , New York, J.Wiley & Sons.

Crawley, M.J. (2007) *The R Book*, Chicester, J.Wiley & Sons.

Crivellari, F. (2006) *Analisi statistica dei dati con R* , Milano, Apogeo.

Daalgard, P. (2002) *Introductory Statistics with R* , New York, Springer.

Espa, G. e Micciolo, R. (2008) *Problemi ed esperimenti di Statistica con R* , Milano, Apogeo.

Faraway, J.J. (2005) *Linear Models with R* , Boca Raton, FL, Chapman & Hall/CRC.

– (2006) *Extending the Linear Models with R* , Boca Raton, FL, Chapman & Hall/CRC.

Fox, J. (2002) *An R and S-PLUS Companion to Applied Regression*, Thousand Oaks, CA, Sage.

Gentleman, R. (2009) *R Programming for Bioinformatics*, Boca Raton, FL, Chapman & Hall/CRC.

Iacus, S. e Masarotto, G. (2003) *Laboratorio Statistico con R* , Milano, McGraw-Hill.

Jones, O., Maillardet, R. e Robinson, A. (2009) *Introduction to Scientific Programming and Simulation Using R* , Boca Raton, FL, Chapman & Hall/CRC.

Maindonald, J. e Braun, J. (2003) *Data Analysis and Graphics Using R . An Example-based Approach*, Cambridge, Cambridge University Press.

Murrell, P. (2005) *R Graphics*, Boca Raton, FL, Chapman & Hall/CRC.

Pauli, F., Torelli, N. e Trevisani, M. (2008) *Statistica: esercizi ed esempi*, Milano, Paravia/Bruno Mondadori.

Quian, S.S. (2010) *Environmental and Ecological Statistics with R* , Boca Raton, FL, Chapman & Hall/CRC.

Rizzo, M.L. (2008) *Statistical Computing with R* , Boca Raton, FL, Chapman & Hall/-CRC.

Stefanini, F.M. (2007) *Introduzione alla Statistica applicata con esempi in R* , Milano, Paravia/Bruno Mondadori.

Sarkar, D. (2008) *Lattice. Multivariate Data Visualization with R* , New York, Springer.

Spector, P. (2008) *Data Manipulation with R* , New York, Springer.

Verzani, J. (2005) *Using R for Introductory Statistics*, Boca Raton, FL, Chapman & Hall/CRC.

Parte prima

Analisi esplorativa delle informazioni

Capitolo secondo

Rilevazioni statistiche

2.1 Introduzione

Una disamina accurata della tipologia delle informazioni disponibili rende più agevole l'applicazione dei metodi statistici; per questo, è importante descrivere accuratamente il fenomeno in esame, il contesto in cui viene studiato ed i limiti (temporali, territoriali o di altra natura) che inevitabilmente condizionano ogni ricerca su dati reali.

In questo capitolo, dopo aver esaminato la natura e le caratteristiche più importanti dei collettivi statistici, presenteremo alcune semplici operazioni che aiutano nella descrizione iniziale dei fenomeni da studiare, anche mediante alcune esemplificazioni su dati reali.

2.2 Elementi di una rilevazione statistica

Spesso i termini statistici coincidono con i vocaboli del linguaggio corrente; ma è necessario precisare con rigore la terminologia scientifica per semplificare l'apprendimento ed applicare correttamente i metodi successivi. Per ciascuna definizione, poi, forniremo alcune esemplificazioni, riassumendo alla fine i vari elementi della rilevazione statistica. Alcune definizioni saranno anche riprese nel Capitolo seguente.

- ***Rilevazione statistica*** è il complesso di operazioni rivolte ad acquisire una o più informazioni su un insieme di elementi oggetto di studio.

Le rilevazioni statistiche possono essere classificate in molti modi:

i) rispetto alla *complessità delle operazioni*: semplici (per esempio, misurare l'altezza in cm di un individuo, chiedere sesso e data di nascita ad un impiegato); oppure complesse (per esempio, codificare un bilancio aziendale; registrare con ritmo orario il prezzo di tutte le azioni contrattate alla Borsa di Milano, valutare il ritmo di accrescimento di cellule tumorali).

ii) rispetto alla *natura delle informazioni* raccolte: risposte (per esempio, opinioni, informazioni personali, gusti, scale di giudizi, sentimenti, etc.) oppure misure (che sono ottenute tramite strumenti: metro, bilancia, orologio, microscopio elettronico, radar, dosatori ormonali, etc.);

iii) rispetto al *gruppo di riferimento*: globali (per esempio, i censimenti, lo studio di tutti i laureati di un certo Ateneo, gli acquisti di tutti i clienti di un supermercato) e parziali (per esempio, i sondaggi di opinione, le interviste telefoniche, i questionari postali).

o Alcune precisazioni aiutano a comprendere meglio le precedenti distinzioni. Anzitutto, si parla di *popolazione* quando il collettivo di riferimento esaurisce tutte le informazioni che si ritengono utili per l'indagine statistica; si parla di *campione* quando tali informazioni sono derivate da un sotto-insieme proprio della popolazione di riferimento. Questo implica che, per esempio, gli studenti presenti in aula in un certo giorno costituiscono la popolazione se l'obiettivo dell'indagine è svolgere un compito, garantire la sicurezza dei presenti in aula, etc., ma essi costituiscono anche un campione di tutti gli studenti iscritti alla Facoltà se obiettivo dell'indagine è stabilire una valutazione sui desideri degli studenti, sui loro orientamenti professionali, etc.

o In passato, si sono accentuati importanza e rigore dei censimenti soprattutto con riferimento a quelli decennali che, quasi ovunque nel mondo, vengono effettuati per registrare lo «stato» della popolazione di una nazione. Nell'ipotesi di una popolazione sostanzialmente stabile sul territorio e ben codificabile, essi possiedono un valore conoscitivo ineliminabile. Tuttavia, negli ultimi decenni, è divenuto prioritario il discorso sulla qualità dei dati raccolti sia per l'accresciuta mobilità della popolazione che per la difficoltà di rendere omogenea la formazione dei rilevatori in tempi molto ridotti.

Per contro, la metodologia statistica ha migliorato la qualità scientifica dei metodi campionari per cui talune informazioni possono (e talora debbono) essere raccolte mediante rilevazioni parziali che addirittura garantiscono una migliore qualità dei dati. Anche se affronteremo tali approcci più oltre, va anticipato che, per sua natura, la rilevazione campionaria rischia sempre di non essere rappresentativa, per cui è necessario porre particolare cautela nella pianificazione di tali rilevazioni ed esaminare i connessi risultati sempre con adeguato senso critico.

o Molte indagini presentate con disinvoltura dai mezzi di comunicazione rischiano costantemente di violare uno o più dei requisiti fondamentali per la correttezza formale e metodologica di una rilevazione statistica. Difatti, la scelta degli elementi da intervistare non risponde quasi mai a criteri di natura statistica perché, per esigenze contingenti, sono derivati dal requisito della disponibilità fisica o cronologica a fornire informazioni: il che evidentemente viola la rappresentatività del collettivo parziale in rapporto alla popolazione.

• ***Popolazione*** (o Universo) è qualsiasi insieme di elementi che forma oggetto di uno studio statistico.

Si distingue tra *popolazione reale* (quando essa è effettivamente esistente e visibile) e *popolazione virtuale* (quando essa non è osservata né osservabile, perché astratta o connessa al futuro, ma è comunque ben definita).

Una popolazione non costituisce necessariamente un insieme biologico. Infatti, sono *popolazioni reali* la popolazione delle lampadine prodotte nell'ultimo mese da un'azienda di Milano, la popolazione delle stelle della Via Lattea, quella dei sostenitori di un prefissato partito politico. Invece, sono *popolazioni virtuali* la popolazione degli acquirenti di un certo modello di automobile che si sta progettando, quella degli studenti che l'anno prossimo supereranno l'esame di Statistica, etc.

• ***Campione*** è qualsiasi sotto-insieme derivato da una certa popolazione e finalizzato ad uno studio statistico.

o La metodologia statistica ha approfondito sia le modalità mediante le quali da una prefissata popolazione si perviene alla individuazione di quel sotto-insieme su cui successivamente si opera (e questo formerà oggetto della *Teoria dei campioni*, che discuteremo per taluni aspetti nel Capitolo 16) che le proprietà delle procedure statistiche applicate ai campioni (e questo formerà oggetto dell'*Inferenza statistica*, di cui si occuperanno le parti terza e quarta di questo volume).

o Tuttavia, va chiarito subito che la dicotomia popolazione-campione si risolve solo in funzione degli obiettivi dell'indagine. Così, i residenti nel comune di Firenze sono certamente un campione degli italiani ma sono anche la popolazione dei residenti a Firenze e l'analisi delle caratteristiche di tali elementi sarà svolta con metodologie differenti a seconda che l'indagine punti a studiare la collettività dei fiorentini oppure quella degli italiani.

o D'altra parte, poiché la conoscenza umana non può che essere parziale e provvisoria, la Statistica privilegia un approccio allo studio dei fenomeni che presuppone sempre una dimensione campionaria, interpretando in ogni caso le informazioni di cui si dispone come un insieme limitato e circoscritto rispetto alla totalità delle informazioni di cui si poteva disporre.

• ***Unità statistica*** (o soggetto) è l'elemento di base della popolazione sul quale viene effettuata la rilevazione di uno o più fenomeni oggetto della indagine.

L'unità statistica può essere ulteriormente suddivisa per le esigenze dell'indagine. Per esempio, la famiglia può essere una unità statistica se si rilevano i componenti ma è composta da varie unità statistiche se si misurano le altezze dei suoi componenti adulti.

È importante sottolineare che l'unità statistica va definita in termini di occasione, tempo, durata, territorio e che, talvolta, essa cambia durante la rilevazione. Per esempio, la popolazione italiana nell'ultimo secolo va riferita ai confini attuali della nazione, chiarendo se essa è calcolata con o senza gli italiani residenti all'estero, se è popolazione presente o popolazione residente, se va riferita ad una data intermedia oppure all'inizio dell'anno solare, e così via. Similmente, uno studio sui prezzi dei sistemi di elaborazione automatica negli ultimi 30 anni non è banale, perché la natura stessa di tali «sistemi» è così cambiata da porre in discussione la utilità di considerare una unica definizione come esauriente per individuare correttamente la popolazione di riferimento.

• ***Variabile*** è il fenomeno oggetto dello studio, rilevato o misurato sulle unità statistiche. Più oltre, distingueremo le variabili in qualitative e quantitative.

Nella letteratura statistica italiana, si parla talvolta di «carattere» anziché di variabile, ma in questo volume utilizzeremo il termine variabile, anche in coerenza con il linguaggio internazionale.

È piuttosto comune che le variabili assumano valori differenti nelle varie unità statistiche e sulla natura di tale diversità si discuterà nel paragrafo seguente.

D'altra parte, sulle popolazioni, e quindi sulle unità statistiche, possono essere rilevate una o molte variabili di interesse. Per esempio, per la popolazione dei residenti in Italia in un certo periodo, si possono rilevare: sesso, religione, età, professione, altezza, reddito mensile, metri quadrati della propria abitazione, stato civile, possesso di titoli di stato, titolo di studio più elevato posseduto, etc.

• ***Modalità*** è l'espressione concreta mediante la quale la variabile si manifesta nelle unità statistiche.

Quindi, modalità è il numero o l'attributo che l'unità statistica manifesta nella rilevazione. L'elenco di tutte le possibili modalità di una variabile si dice *esaustivo* se tale elenco è completo; le modalità si dicono *esclusive* (o *disgiunte*) se una unità statistica può manifestare la variabile in una ed una sola modalità tra quelle indicate.

Esempio 2.1

o La registrazione «Adriano è alto cm 173» costituisce una rilevazione sulla *unità statistica* Adriano per il quale la *variabile qualitativa* «Sesso» si manifesta nella *modalità* (=attributo) Maschio e la *variabile quantitativa* «Altezza» si manifesta nella *modalità* (=numero) 173 cm.

o In una indagine sulla età degli studenti iscritti al 1° anno della Facoltà di Scienze Politiche dell'Università di Napoli Federico II, l'informazione statistica contenuta nella frase «Antonio dichiara di aver compiuto 19 anni» può essere così schematizzata:

i) la *popolazione* oggetto dello studio è costituita dall'insieme degli studenti sopra definiti;

ii) l'*unità statistica* è Antonio perché è uno degli elementi della popolazione in esame;

iii) la *variabile* che si sta esaminando è l'Età che – se misurata in anni compiuti – nella popolazione in esame può assumere una delle seguenti modalità $18, 19, 20, \ldots$, che sono disgiunte ed esaustive;

iv) il valore 19 è la specifica *modalità* della variabile «Età» registrata sull'unità Antonio.

o In una indagine circa le preferenze della marca di automobili, in funzione del sesso e della regione di residenza, la rilevazione ha prodotto il risultato: «Maria Rossi risiede nel Lazio e possiede una Lancia-δ». Allora, stiamo esaminando la popolazione dei maggiorenni italiani possessori di autoveicoli, Maria Rossi è l'unità statistica prescelta sulla quale sono rilevate le tre variabili: Sesso, Regione di residenza, Marca di autoveicolo posseduto, le quali hanno assunto gli attributi: {Femmina, Lazio, Lancia-δ}.

□

L'esempio mostra che una popolazione si specifica nella unità statistica mentre la variabile (che varia nella popolazione) si specifica nella modalità (assunta nella unità statistica).

Si osservi che ogni studio – per essere concretamente utilizzabile – delimita le informazioni statistiche da rilevare, sino al punto da considerare *equivalenti* due soggetti che possiedono uguali le sole caratteristiche oggetto della indagine medesima. In altri termini, due unità statistiche possono essere molto diverse

tra loro ma possono essere considerate due repliche della stessa informazione se assumono le stesse modalità per le stesse variabili oggetto di studio.

o Se obiettivo dell'indagine è lo studio dell'altezza in rapporto al sesso, Maria e Giovanna – entrambe alte cm 168 – sono considerate indistinguibili, anche se abitano a 600 km di distanza, possiedono redditi molto differenti, la prima è una ventenne e nubile, mentre la seconda è una quarantenne e sposata con figli, una è diplomata e l'altra è laureata, e così via.

• ***Frequenza*** è il numero di volte in cui una determinata modalità si verifica nel collettivo di riferimento. La frequenza è l'elemento caratterizzante le distribuzioni di frequenza cui sarà dedicato il Capitolo 3.

• ***Serie*** è l'insieme delle modalità rilevate sulla popolazione di riferimento ed organizzate in modo che a ciascuna unità della popolazione corrisponde una ben definita modalità. La sequenza di tali modalità può essere organizzata secondo criteri molto diversi: geografico, alfabetico, cronologico, cromatico, casuale, etc.

• ***Seriazione*** è l'insieme delle modalità di una popolazione di riferimento organizzate in modo che a ciascuna modalità corrisponda la rispettiva frequenza. Ai nostri fini, l'insieme dei dati disponibili mediante una seriazione definisce una *distribuzione di frequenza*, come studieremo in dettaglio a partire dal Capitolo seguente.

o In pratica, quando si dispone di rilevazioni sulle singole unità statistiche isolatamente considerate si ha una *serie* (territoriale, temporale, etc.). Quando, invece, queste informazioni sono organizzate in tabelle nelle quali alle modalità della variabile corrispondono le rispettive frequenze si è in presenza di una *seriazione*. Talvolta, nel seguito, chiameremo *distinti* i dati disponibili come serie, e *raggruppati* rispetto alle modalità quelli disponibili come seriazione.

Esempio 2.2

o Supponiamo che in un'aula vi siano 25 studenti per i quali si registra la variabile (qualitativa) «Sesso» e che la rilevazione abbia fornito i seguenti risultati:

{ M F F F M F M F M M F F F F M F F M F M F M F M F }

La precedente elencazione costituisce la *serie* della variabile «Sesso» nella popolazione in esame.

Se, invece, organizziamo gli stessi dati in una tabella ove facciamo corrispondere alle due possibili modalità (M, F) le corrispondenti frequenze abbiamo allora una *seriazione* (detta anche distribuzione di frequenza).

Modalità	*Frequenze*
Femmine	15
Maschi	10
Totale	25

□

2.3 Tipologia delle informazioni statistiche

Dal paragrafo precedente è emerso che ogni risultato ottenuto da una indagine sui collettivi esaminati (popolazione o campione) rispetto ai loro costituenti (=unità statistiche) e in rapporto ad uno o più fenomeni (=variabili) costituisce una informazione statistica. Dobbiamo ora distinguere con maggiore dettaglio le differenti informazioni che si rendono disponibili, al fine di proporre le metodologie più convenienti.

La classificazione delle variabili è un argomento molto complesso che ha coinvolto sociologi, psicologi, statistici e matematici di varie generazioni. La letteratura, quindi, propone numerose distinzioni con molti altri dettagli; inoltre, talune soluzioni cui si perviene sono funzionali ai particolari settori scientifici in cui si opera. Per questo, semplificando tra le tante scelte possibili, abbiamo proposto in questo testo una suddivisione essenziale e che giustifica i differenti indicatori statistici che via via si proporranno.

Un fenomeno si definisce ***variabile*** se assume per modalità dei numeri reali, mentre un fenomeno si definisce ***mutabile*** oppure ***variabile qualitativa*** se assume attributi non-numerici, che possono essere della più varia natura.

Nel seguito, parleremo di ***variabili quantitative*** se le modalità sono numeri e di ***variabili qualitative*** in tutti gli altri casi.

Esempio 2.3

o L'età, il peso, l'altezza, la temperatura, il voto all'esame, il tempo impiegato per raggiungere l'aeroporto, qualsiasi grandezza fisica misurabile, la proporzione di favorevoli ad una legge, il numero di clienti in attesa ad uno sportello bancario, il tasso di accrescimento di una coltura batterica, i nati vivi in provincia di Rieti, il numero dei passeggeri annui che transitano per l'aeroporto di Roma-Fiumicino, etc. costituiscono esempi di fenomeni di natura quantitativa, quindi di *variabili*.

o Il sesso, la religione, lo stato civile, la professione, la categoria sociale di appartenenza, il comune di nascita, la regione di residenza, la nazionalità, lo stato d'animo dopo un esame, il gusto, il gradimento di una canzone, etc. costituiscono esempi di fenomeni di natura qualitativa, quindi di *mutabili* (=variabili qualitative). Per tali variabili, ogni associazione con numeri reali – se anche è possibile o talvolta conveniente, per finalità di codifica – è sostanzialmente arbitraria.

□

Le *variabili qualitative* si distinguono in nominali e ordinali; le *variabili quantitative* si distinguono in discrete e continue, secondo lo schema e le definizioni seguenti.

Variabili	*Nominali*
QUALITATIVE	*Ordinali*

Variabili	*Discrete*
QUANTITATIVE	*Continue*

• Una ***variabile nominale*** è una variabile qualitativa le cui modalità (attributi) non assumono alcun ordine precostituito. L'unico confronto ammissibile tra due unità statistiche rispetto alle modalità assunte da variabili nominali consiste nello stabilire se possiedono o no lo stesso attributo, cioè se sono diversi oppure uguali rispetto a quella variabile.

o Il sesso, la religione, la professione, la nazionalità delle persone costituiscono esempi di variabili nominali i cui attributi (Maschi e Femmine; Cristiani, Ebrei, Mussulmani, Induisti, Altri; Lavoratore dipendente, Indipendente, Pensionato; Italiana, Francese, Tedesca, etc.) non possiedono un ordine naturale né esistono motivi logico-concettuali per preferire un ordinamento tra le modalità. Quindi, il confronto tra due unità statistiche rispetto al sesso, alla religione, alla professione, alla nazionalità, etc. permette solo di decidere se le due unità possiedono la stessa modalità oppure no (cioè se sono dello stesso sesso, della stessa religione, della stessa professione, della stessa nazionalità, etc.).

• Una ***variabile ordinale*** è una variabile qualitativa le cui modalità sono logicamente sequenziali, in ordine crescente o decrescente. Quindi, per tali variabili, pur essendo impossibile effettuare operazioni aritmetiche tra gli attributi delle singole unità statistiche, si può stabilire tra di esse una *relazione di ordine* rispetto alla variabile individuata.

o Il grado militare (sergente, tenente, capitano, maggiore, colonnello, generale), la valutazione scolastica (insufficiente, mediocre, sufficiente, buono, ottimo), un indice di gradimento (per niente, poco, abbastanza, molto) sono fenomeni nei quali la gerarchia, il merito, la preferenza individuale stabiliscono una graduatoria che mostra univocamente un criterio di ordinamento (precedenza) tra le unità statistiche, anche se non è possibile attribuire un valore numerico alla distanza tra le modalità.

• Una ***variabile discreta*** è una variabile quantitativa se le modalità che può assumere possono essere poste in corrispondenza con l'insieme dei numeri naturali $\{1,2,...\}$ o un suo sottoinsieme proprio. In sostanza, le variabili discrete assumono un insieme numerico di modalità *elencabili*.

o Il numero di clienti in attesa ad uno sportello bancario, i nati vivi in provincia di Rieti, il numero dei lanci di un dado necessari perché si verifichi un punto «5» per la prima volta, il voto all'esame di Statistica, il numero dei componenti di una famiglia, etc. costituiscono esempi di variabili discrete perché *concettualmente* le possibili modalità di tali variabili sono elencabili e, quindi, possono essere messe in corrispondenza con i numeri naturali.

Infatti, il voto ad un esame universitario può assumere solo le modalità contenute in $\{18, 19, \ldots, 29, 30\}$; il numero dei clienti in fila di attesa può assumere valori in $\{0,1,2, \ldots, M\}$ essendo M un intero massimo predefinito in funzione dei potenziali utenti; il numero dei lanci di un dado necessari perché si verifichi «5» per la prima volta può assumere modalità contenute in $\{1,2,3, \ldots\}$ ed è quindi una variabile discreta con un numero infinito ma numerabile di modalità, etc.

• Una ***variabile continua*** è una variabile quantitativa le cui modalità possono assumere, in linea di principio, qualsiasi valore contenuto in un intervallo reale.

Si osservi che una variabile è continua *sul piano concettuale* in quanto, se si considera ciò che avviene in pratica, ogni indagine sarà costretta ad approssimare le misurazioni dei fenomeni reali, e quindi di fatto a rendere discrete quelle variabili che per definizione sono continue.

o L'età, il peso, l'altezza, la temperatura di una stanza, il tempo impiegato per raggiungere l'aeroporto, qualsiasi grandezza fisica misurabile, etc. costituiscono esempi di variabili continue poiché le modalità di tali variabili possono assumere qualsiasi valore in un intervallo predefinito; né esistono motivi concettuali per cui, per esempio, fissati due pesi estremamente vicini, non sia possibile trovare una persona con peso intermedio a quei pesi.

Ovviamente, in pratica, la nostra possibilità di valutare punti estremamente vicini dipende dagli strumenti di misura che consentono misure accurate solo fino ad un certo punto. Così nelle altezze ci si limita al massimo a considerare i millimetri, nelle gare di Formula 1 la velocità delle automobili viene misurata in km/h sino a tre decimali, nella determinazione della età ci si accontenta generalmente degli anni compiuti, etc. Tutte queste operazioni costituiscono delle *discretizzazioni* di variabili che concettualmente sono, invece, continue.

Generalmente, una singola variabile si indica con la lettera X e le modalità che essa assume con $x_1, x_2, \ldots$ Quando le variabili sono più di una si parla di variabili doppie, triple, ..., multiple e le si indica, rispettivamente, con (X,Y); (X,Y,Z); ..., $(X,Y,\ldots,W)$. Le modalità che esse assumono sulle singole unità statistiche saranno, rispettivamente:

- coppie ordinate: $(x_1,y_1),(x_2,y_2),\ldots$
- triple ordinate: $(x_1,y_1,z_1),(x_2,y_2,z_2),\ldots$
- m-ple ordinate: $(x_1,y_1,\ldots,w_1),(x_2,y_2,\ldots,w_2),\ldots$

Anche per le informazioni statistiche multiple, si parla di *serie* quando le informazioni vengono presentate distinte per ciascuna unità statistica e di *seriazione* quando vengono raccolte per modalità e corrispondenti frequenze. In tale ambito, anche grazie alla diffusione dei fogli elettronici, è molto diffusa la presentazione nelle cosiddette «matrici di dati».

La ***matrice di dati*** è una rappresentazione tabellare mediante la quale si schematizzano le informazioni (misure, registrazioni o risposte) raccolte su ciascuna unità statistica in rapporto ad una molteplicità di variabili. In sostanza, la matrice dei dati costituisce la serie dei dati su una variabile statistica multipla.

Ogni *colonna* della matrice esprime una variabile (qualitativa o quantitativa) rilevata sulle diverse unità statistiche.

Invece, ogni *riga* della matrice esprime ordinatamente le rilevazioni su tutte le variabili ottenute per una singola unità statistica.

Esempio 2.4

o Il questionario rivolto agli studenti frequentanti il corso di Statistica contiene numerose domande di natura qualitativa e quantitativa. Le risposte fornite da ciascuno studente costituiscono una riga della matrice dei dati; invece, ogni colonna rappresenta le informazioni sulla specifica variabile fornite da tutti gli studenti.

o Le rappresentazioni tabellari pubblicate sui giornali sportivi che, in corrispondenza di ciascuna squadra di calcio, presentano i punti, il numero di gol fatti e subiti, le partite vinte, i pareggi e le sconfitte, sono una matrice dei dati.

o Una differente matrice dei dati è una tabella nella quale ciascuna riga rappresenta una nazione e ciascuna colonna esprime il valore di un indicatore socio-economico o culturale misurato su ciascuna nazione, come viene discusso nell'esempio seguente. □

Esempio 2.5

o Presentiamo qui una matrice di dati contenente alcuni indicatori socio-demografici, per gli anni 1990-93, relativi a 20 nazioni tra loro molto diverse sul piano del benessere e della diffusione delle risorse economiche disponibili. I dati sono di fonte ONU e contengono un dato mancante, indicato con (.), per la variabile X_7 sulla unità statistica «Russia».

Le variabili considerate sono le seguenti:

X_1 = Consumi elettricità kWh/pro-capite
X_2 = Persone per veicolo
X_3 = Persone per apparecchio telefonico
X_4 = Percentuale della popolazione con acqua potabile
X_5 = Speranza di vita dei maschi, in anni
X_6 = Mortalità infantile
X_7 = Percentuale di donne alfabete oltre i 15 anni
X_8 = Tasso di natalità per 1000 abitanti.

Ogni riga costituisce il profilo di una Nazione (=unità statistica) ed ogni colonna costituisce la distribuzione (=serie) della variabile indicata rispetto alle 20 unità statistiche. Nella tabella seguente, le nazioni sono elencate in ordine decrescente rispetto alla variabile X_1.

Nazioni	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Stati Uniti	12281	1.3	1.3	100	72.0	8.5	95.3	15.9
Australia	9045	1.86	2.1	99	73.9	8.2	99.5	15.1
Germania	7199	2.1	1.7	100	72.2	7.5	100.0	11.4
Giappone	7161	2.1	1.8	100	75.9	4.6	100.0	9.9
Russia	5818	16.0	6.1	100	63.5	17.8	.	12.1
Italia	4452	1.9	1.8	100	73.5	8.3	96.4	9.8
Portogallo	3037	4.0	3.2	92	70.6	11.0	86.9	11.8
Brasile	1723	11.0	10.0	97	63.5	57.0	80.6	26.1
Argentina	1601	5.7	7.0	65	68.0	31.0	95.1	20.0
Malesia	1545	7.6	8.9	79	69.0	13.0	70.4	27.0
Messico	1440	8.5	8.0	71	66.5	24.2	85.1	31.2
Colombia	1116	24.0	8.7	88	66.4	37.0	85.9	25.8
Egitto	754	38.0	22.0	73	59.0	49.6	33.8	30.8
Cina	592	87.0	77.0	74	68.6	33.0	68.0	18.2
India	360	220.0	145.0	86	57.7	80.0	33.7	32.5
Indonesia	248	44.0	122.0	58	55.6	90.0	70.0	32.5
Kenya	138	76.0	66.0	30	59.0	63.0	49.2	47.0
Nepal	43	2259.0	275.0	37	55.4	98.0	23.3	39.0
Etiopia	19	984.0	316.0	19	45.4	122.0	0.5	48.4
Cambogia	8	737.0	3300.0	18	46.5	131.0	48.0	41.8

□

2.4 Operazioni statistiche elementari

Alcune operazioni elementari consentono un primo esame dei dati statistici. Tra queste ricordiamo le differenze tra le modalità di due variabili quantitative, rilevate in occasioni (tempi, luoghi, unità statistiche, circostanze, etc.) differenti.

La ***differenza (assoluta)*** tra due modalità x_1 e x_2 di una variabile quantitativa X è definita da:

$$differenza = x_2 - x_1;$$

mentre la ***differenza relativa*** è pari alla differenza assoluta rapportata ad una delle due modalità (generalmente la prima):

$$differenza\ relativa = \frac{x_2 - x_1}{x_1};$$

infine, la ***differenza percentuale*** è la differenza relativa moltiplicata per 100, cioè:

$$differenza\ percentuale = \frac{x_2 - x_1}{x_1} \times 100\,.$$

Esempio 2.6

∘ Alla *differenza* (assoluta) di cm 4.5 tra l'altezza di cm 174.5 di Marco e quella di cm 170.0 di Maria (le circostanze sono le due unità statistiche) corrisponde una *differenza relativa* (rispetto all'altezza di Maria) che vale $(174.5 - 170.0)/170.0 = 0.0264$, e una *differenza percentuale* di 2.64. Quindi, Marco è più alto di Maria del 2.64%.

∘ Se la velocità massima di una utilitaria è di 120 km/h mentre quella di un'auto sportiva è di 240 km/h, possiamo dire che vi è una *differenza assoluta* in velocità massima di -120 km/h (tra la prima e la seconda auto) e una *differenza relativa* di -1 $(= (120 - 240)/120)$, oppure una *differenza percentuale* di -100%. In sostanza, risulta che l'utilitaria esaminata ha una velocità esattamente pari alla metà dell'auto sportiva. □

La *differenza assoluta* è utile per il confronto tra fenomeni simili in circostanze differenti (perché è espressa nella stessa unità di misura dei fenomeni che si confrontano) mentre la *differenza relativa* è necessaria per confrontare le variazioni di fenomeni espressi in unità di misure differenti (infatti, essa non dipende dall'unità di misura: si dice, talvolta, che è un numero puro). La *differenza percentuale* è solo un modo di effettuare i confronti relativi rispetto al numero 100.

Quando si osserva una variabile X rispetto al tempo, la differenza relativa tra la modalità (x_{t_2}) assunta al tempo t_2 rispetto a quella (x_{t_1}) del tempo t_1, con $t_1 < t_2$, viene definita ***tasso di variazione*** (percentuale, se il risultato è moltiplicato per 100) di X nel periodo $[t_1, t_2]$:

$$\frac{x_{t_2} - x_{t_1}}{x_{t_1}} \times 100\,.$$

∘ Se la variabile X assume valori strettamente positivi (prezzi, salari, investimenti, peso, durata, altezza, velocità, etc.), allora la differenza tra i logaritmi (neperiani) delle modalità rilevate in due tempi distinti equivale, con buona approssimazione, al tasso di variazione.

Infatti, è immediato dimostrare che:

$$log(x_{t_2}) - log(x_{t_1}) = log(\frac{x_{t_2}}{x_{t_1}}) = log\left(\frac{x_{t_1} + x_{t_2} - x_{t_1}}{x_{t_1}}\right)$$
$$= log\left(1 + \frac{x_{t_2} - x_{t_1}}{x_{t_1}}\right) \simeq \frac{x_{t_2} - x_{t_1}}{x_{t_1}},$$

perché, per piccoli valori di z, risulta: $log(1 + z) \simeq z$. Ne consegue che l'approssimazione logaritmica è accettabile quando la variazione relativa è molto piccola.

Esempio 2.7

∘ Per esempio, dire che l'indice MIBTEL è variato oggi di +0.84% significa che le quotazioni medie rilevate da questo indicatore nella Borsa di Milano sono cresciute, rispetto a ieri, dello 0.84%, cioè il tasso di variazione è stato pari a +0.0084.

∘ Similmente, affermare che l'inflazione mensile (detta anche *inflazione congiunturale*) è stata in Gennaio dello 0.4% significa che l'indice generale dei prezzi al consumo, pubblicato dall'ISTAT, è aumentato, rispetto al mese precedente, dello 0.4%. Quindi, il tasso di variazione è stato di +0.0040.

∘ Ancora, quando si afferma che l'*inflazione tendenziale* in Novembre 2003 è stata del 2.6% significa che l'indice generale dei prezzi al consumo di Novembre 2003 rispetto all'analogo indice calcolato dall'ISTAT nel Novembre 2002 è cresciuto del 2.6%. Quindi, nel periodo indicato, il tasso di variazione è stato di +0.026.

∘ A tale riguardo, va sottolineato come l'affermazione: «l'inflazione è ferma» non significa che i prezzi sono costanti ma semplicemente che *la crescita dei prezzi avviene con ritmo costante* nell'intervallo considerato. Allo stesso modo, dire che «la velocità di un'auto è costante» non implica che essa è ferma ma solo che le sue variazioni spaziali nell'unità di tempo sono costanti, cioè che si muove con velocità costante. □

2.5 Rapporti statistici

In questo paragrafo presenteremo alcuni indici elementari (molto semplici e di uso frequente) che vengono definiti *rapporti statistici.*

• Il ***rapporto di composizione*** si ottiene dividendo il valore rilevato in una certa circostanza per l'analogo valore rilevato per l'intera popolazione (e, se opportuno, moltiplicando tale rapporto per 100).

Esso esprime, quindi, la frazione relativa (o percentuale) posseduta o registrata dalla modalità in questione rispetto al totale.

Esempio 2.8

∘ La composizione percentuale della forza lavoro rispetto al sesso, alle fasce di età, al titolo di studio più elevato posseduto, etc. costituiscono esempi di rapporti di com-

posizione. Similmente, la distribuzione relativa della popolazione residente tra Giovani, Attivi, Anziani, oppure quella della tipologia della clientela di una banca, oppure quella dei componenti elementari di una sostanza fisica, oppure il rapporto tra i morti ed il totale dei passeggeri (morti+feriti) coinvolti in incidenti stradali (detto *indice di gravità* degli incidenti stradali), etc. sono ulteriori esempi di rapporti di composizione.

□

• Il ***rapporto di derivazione*** si ottiene dividendo la modalità di una variabile per quella corrispondente di un'altra che, sul piano logico e/o temporale, ne costituisce causa o ne è antecedente necessario.

Per tali indici, è importante stabilire correttamente il collettivo di riferimento più idoneo, ponendo al denominatore la «effettiva» popolazione che può aver generato il dato collocato al numeratore, e non un collettivo di riferimento generico. Per questo, si distingue tra *rapporti specifici* e *rapporti generici* di derivazione. Quando il collettivo di riferimento al denominatore è così ristretto che esso, e solo esso, può generare il collettivo al numeratore si parla di *rapporto di causalità.*

Esempio 2.9

o Alcuni rapporti di derivazione ricevono nomi specifici. Per esempio, in Demografia: *indice di natalità, indice di mortalità, tasso di mortalità infantile*; nelle Scienze economiche: *indici di produttività*; nelle Scienze sociali: *indici di criminalità*, etc.

o Il numero dei nati e dei morti nelle regioni italiane è influenzato dalla corrispondente popolazione residente, per cui è più corretto confrontare i rapporti (nati/popolazione) e (morti/popolazione), ove la popolazione viene determinata convenzionalmente a fine anno, oppure come semisomma di quella residente all'inizio e alla fine dell'anno.

Per quanto concerne il rapporto nati/popolazione, bisognerebbe per coerenza porre al denominatore le sole donne in età feconda stabilmente presenti nel territorio prefissato nei nove mesi precedenti il periodo cui fanno riferimento le nascite.

o Il numero di laureati di una Facoltà dipende dal numero degli iscritti, per cui uno studio comparato sulla «propensione a terminare gli studi» va correttamente impostato tenendo conto del rapporto (laureati per anno)/(immatricolati negli ultimi anni), oppure seguire il percorso di un gruppo di studenti immatricolati in un certo anno sino alla fine degli studi (si parla in tal caso di *coorte*).

o Negli studi statistici sul traffico stradale, si parla di *indice di pericolosità* come rapporto tra numero di morti in incidenti stradali e numero di sinistri che li hanno causati, in un certo periodo di tempo. Se esso è analizzato rispetto al tempo, al territorio, alla tipologia delle strade, all'infrazione che è stata all'origine dell'incidente permette di quantificare l'efficacia di norme restrittive e di interventi rivolti alla sicurezza dell'autoveicolo e/o della rete stradale.

□

• Il ***rapporto di densità*** è definito mediante il confronto tra la dimensione globale di un fenomeno (al numeratore) e la dimensione spaziale, temporale o caratterizzante cui esso fa riferimento (al denominatore). Talvolta, l'inverso del rapporto di densità si chiama *rapporto di estensione.*

Esempio 2.10

o Tra i rapporti di densità più comuni ricordiamo la *densità della popolazione* residente rispetto alle regioni o altri ambiti territoriali, il numero di componenti per famiglia, il numero di persone per vano di una abitazione (detto anche «*quoziente di affollamento*»), la densità di una sostanza inquinante nell'atmosfera, il numero di veicoli circolanti rispetto alla rete autostradale, e così via.

In alcuni di tali esempi, le misure inverse (cioè i corrispondenti *rapporti di estensione*) vengono effettivamente utilizzati: per esempio, si parla di quanti m^2 di verde dispone ogni residente in una città.

□

• Il ***rapporto di coesistenza*** riguarda il rapporto tra la frequenza (o quantità) di una modalità rispetto alla frequenza (o quantità) corrispondente di un'altra modalità.

Esempio 2.11

o I più comuni rapporti di coesistenza sono utilizzati negli studi demografici e, per il rilievo che hanno per le implicazioni socio-politiche ed economiche, assumono una terminologia specifica:

i) il *rapporto di mascolinità* è definito come il rapporto tra la frequenza della modalità «Maschi» e quella della modalità «Femmine», riferito ad un certo luogo, tempo o circostanze (per esempio, alla nascita); se moltiplicato per 100, esso misura quanti Maschi vi sono per 100 Femmine. L'inverso di tale rapporto (eventualmente moltiplicato per 100) è il *rapporto di femminilità*;

ii) l'*indice di vecchiaia* è definito come il rapporto tra i residenti con più di 65 anni e quelli con meno di 15 anni; se moltiplicato per 100, esso misura quanti anziani vi sono per 100 giovani;

iii) l'*indice di dipendenza degli anziani* è definito come il rapporto tra i residenti con età superiore a 65 anni e quelli di età compresa tra 15 e 65 anni; se moltiplicato per 100, esso misura quanti anziani sono economicamente sostenuti da 100 individui attivi.

□

• L'***indice di eccedenza*** di una variabile che assume due sole modalità misura di quanto la frequenza di una modalità supera l'altra in rapporto all'ammontare totale del fenomeno.

Se, quindi, alle due sole modalità x_1, x_2 di una variabile X corrispondono le frequenze (assolute) n_1, n_2, allora l'indice di eccedenza della prima modalità sulla seconda è definito da:

$$\textit{Indice di eccedenza} \;=\; \frac{n_1 - n_2}{n_1 + n_2}.$$

È agevole convincersi che gli indici di eccedenza variano tra -1 e $+1$ (tra -100 e $+100$, se vengono moltiplicati per 100) e sono utili per confronti specifici (come nell'esempio seguente). Tuttavia, essi non aggiungono nuove informazioni statistiche rispetto ai *rapporti di coesistenza* con i quali sono strettamente connessi

dalle formule seguenti:

$$Indice\ di\ eccedenza\ =\ \frac{1 - Rapporto\ di\ coesistenza}{1 + Rapporto\ \ di\ coesistenza}.$$

e, viceversa,

$$Rapporto\ di\ coesistenza = \frac{1 - Indice\ di\ eccedenza}{1 + Indice\ di\ eccedenza}.$$

Esempio 2.12

∘ L'indice di eccedenza può essere utilizzato, per esempio, per tener conto degli studenti che entrano in una Facoltà per immatricolazione e quelli che ne escono perché laureati; per misurare lo squilibrio del commercio internazionale calcolando le differenze tra importazioni ed esportazioni complessive di un anno ed il volume complessivo degli scambi internazionali (importazioni+esportazioni), etc.

∘ Negli studi demografici i rapporti di eccedenza si applicano per misurare lo squilibrio dei sessi in una popolazione, cioè l'eccedenza di Femmine sui Maschi rispetto al totale della popolazione.

Per esempio, al Censimento del 1981 risultarono 29.051 milioni di Femmine e 27.506 milioni di Maschi, per cui l'indice di eccedenza risulta essere: $\frac{29.051 - 27.506}{29.051 + 27.506} = 0.0273$, cioè, alla data indicata, le donne superavano i maschi del 2.73%.

Tale risultato poteva anche dedursi utilizzando il rapporto di coesistenza (pari in tal caso a: $\frac{27.506}{29.051} = 0.9466$). Infatti: $\frac{1 - 0.9466}{1 + 0.9466} = 0.0273$.

□

• Tra i numerosi altri rapporti statistici, giova ricordare i *rapporti di durata* e i *rapporti di ripetizione*, nonché i ***numeri indice***, sui quali ci soffermeremo con maggiore ampiezza nel paragrafo 4.5.

2.6 Presentazione delle rilevazioni statistiche

Rispetto alla mera elencazione dei risultati di una indagine, è preferibile presentare le informazioni statistiche in una forma organizzata per agevolare i confronti e le analisi successive.

In generale, i dati statistici si possono presentare in forma:

i) *enumerativa*, quando le informazioni statistiche sono disponibili secondo la sequenza delle modalità rilevate sulle singole unità statistiche (è il caso delle serie);

ii) *tabellare*, quando le informazioni statistiche sono presentate secondo un raggruppamento di modalità evidenziato in tabelle, semplici o complesse (è il caso delle seriazioni);

iii) *grafica*, quando le informazioni statistiche sono presentate mediante disegni e schemi grafici, di natura diversa, in funzione della disponibilità dei dati e delle finalità dell'indagine.

Anche se bisogna esercitare notevole senso critico e costante cautela di fronte alle innumerevoli rappresentazioni grafiche diffuse dai mass-media (spesso, con la pretesa di offrire dati ed informazioni statistiche di qualità), va sottolineato come la rappresentazione grafica sia un momento importante per le analisi statistiche. In effetti, il cervello umano possiede, tramite il senso della vista, una velocissima, efficiente e duratura capacità di immagazzinare e confrontare informazioni anche complesse.

Con riferimento alle differenti tipologie, si possono distinguere:

- grafici di *dati statistici rilevati* e delle loro trasformazioni elementari;
- grafici derivati da *risultati di metodi statistici* applicati ai dati rilevati;
- grafici derivati dai *modelli costruiti per i dati* e connessi alla loro verifica, previsione, simulazione, controllo, etc.

Rinviando alla bibliografia per ogni utile approfondimento, ci limitiamo qui a segnalare pochi ma essenziali criteri generali per una corretta rappresentazione grafica.

Per realizzare tali grafici nell'ambiente R , si vedano i paragrafi R 2.9-2.11.

∘ In primo luogo, per ogni grafico occorre esplicitare la natura del fenomeno rappresentato e la fonte delle informazioni, l'unità di misura, la definizione degli assi di riferimento (cartesiani, polari, altri) assieme ai criteri che hanno determinato la costruzione del grafico e le scelte conseguenti.

∘ Poiché lo schermo dei PC è rettangolare, non bisogna sottovalutare la distorsione sistematica che ciò provoca nella rappresentazione grafica quando il software applica una scelta automatica delle scale sugli assi di riferimento.

∘ Per diagrammi cartesiani, la scelta della scala delle ascisse e delle ordinate è un elemento molto delicato poiché il restringimento (allargamento) di una delle due può modificare l'impatto visivo del fenomeno. La questione non può essere affrontata con regole sempre valide ma, in ogni caso, è saggio verificare l'effetto della visualizzazione con una molteplicità di scale.

∘ Occorre evitare le rappresentazioni grafiche che utilizzano figure di fantasia strettamente connesse al fenomeno (ideogrammi, cartogrammi, etc.) perché, pur risultando efficaci per un pubblico non specialistico, queste non rendono quasi mai i rapporti e le differenze relative nelle scale originarie dei fenomeni.

∘ Anche se la sovrapposizione di più grafici omogenei aiuta nella comparazione, una molteplicità di grafici nella stessa figura (spesso con scale differenti) non agevola nella interpretazione dei risultati.

Esempio 2.13

∘ Dal data set STUDENTI, si può esaminare la serie di 2759 modalità:

{CL, CL, SC, SC, TE, TE, TE, CL, AL, SC, AL, AL, CL, SC, SC, TE,..., TE}

riguardanti il Diploma secondario degli studenti (nelle modalità CL=Classico, SC=Scientifico, TE=Tecnico-professionale, AL=Altro). Tuttavia, è sicuramente più immediato ed utile esaminare le stesse informazioni organizzate rispetto alle quattro possibili modalità.

Più semplicemente, sul piano grafico, si può fare riferimento alla Figura R 2.4 (detta grafico «*a torta*») nella quale i settori circolari sono suddivisi in proporzione alla frequenza delle singole modalità.

Per esempio, il settore circolare corrispondente alla modalità «Classico» occupa una frazione pari a 0.2976 (cioè il 29.76%) dell'area complessiva del cerchio ed il corrispondente settore circolare possiede ampiezza angolare al centro pari a circa 107° (valore ottenuto moltiplicando per 0.2976 i 360° dell'angolo giro).

□

• La rappresentazione di gran lunga più importante è quella delle ***distribuzioni di frequenza*** le quali indicano come le unità della popolazione si distribuiscano rispetto alle modalità della variabile in esame.

• Le ***serie storiche*** (dette anche «serie temporali») rappresentano la dinamica di un certo fenomeno registrato istantaneamente o conteggiato durante un periodo prefissato. Inoltre, si parla di serie univariate, bivariate, ..., multivariate a seconda che i fenomeni oggetto di indagine siano $1, 2, \ldots$, molti.

◦ Per le serie storiche la rappresentazione grafica più naturale (e frequente) pone in ascissa il tempo ed in ordinata la variabile misurata per i corrispondenti periodi; quando si hanno più fenomeni contemporanei è comune sovrapporre tali andamenti per visualizzarne le differenze. Ulteriori dettagli sono riportati nel paragrafo R 2.13.

Esempio 2.14

◦ La temperatura del mare misurata ogni mezz'ora, la popolazione residente in Italia censita ogni 10 anni, l'umidità dell'aria misurata ogni ora da uno strumento automatico, etc. costituiscono esempi di serie storiche nelle quali la registrazione della variabile avviene ad un istante prefissato.

◦ I nati vivi registrati ogni anno a Roma, la produzione mensile di una industria chimica, il consumo trimestrale di acqua potabile in Sicilia, il numero di conversazioni telefoniche misurate ogni giorno in partenza da un certo telefono, etc. costituiscono esempi di serie storiche per le quali la registrazione della variabile fa riferimento ad un prefissato intervallo temporale.

◦ Nella tabella seguente presentiamo, per gli anni 1990-2003, la serie mensile dell'indice generale della produzione industriale in Italia (corretta per tenere conto che nei mesi dell'anno i giorni lavorativi sono differenti): essa è reperibile, periodicamente aggiornata, sul sito dell'ISTAT (*www.istat.it*). Il suo andamento temporale è presentato nella Figura R 2.8 (in alto a sinistra).

La natura ripetitiva del fenomeno esaminato rispetto ai mesi dell'anno evidenzia nell'indice della produzione industriale una chiara componente stagionale, caratterizzata dalla quasi totale interruzione dell'attività produttiva per ferie estive. Questo aspetto è già evidente dalla Figura R 2.8 (in alto a sinistra) ma può essere messo in evidenza mediante rappresentazioni grafiche alternative, come la sovrapposizione del profilo annuale della

Indice generale della produzione industriale mensile in Italia 1990-2003 (fonte: ISTAT)

Anni	Gennaio	Febbraio	Marzo	Aprile	Maggio	Giugno	Luglio	Agosto	Settembre	Ottobre	Novembre	Dicembre
1990	86.1	87.3	96.3	89.3	93.8	91.3	93.2	43.4	92.1	94.5	89.8	82.8
1991	87.5	85.5	93.3	86.1	92.3	91.7	92.1	42.0	92.2	93.6	93.0	80.8
1992	85.7	87.4	95.4	87.8	93.1	91.9	92.2	40.1	87.0	91.6	90.9	76.9
1993	83.8	84.7	92.4	84.6	90.4	88.4	88.8	38.7	87.4	89.9	89.3	79.0
1994	82.1	85.4	94.1	89.1	94.6	94.1	95.8	44.7	94.6	96.2	95.8	89.4
1995	91.1	91.8	101.6	92.5	99.9	98.1	99.2	50.9	98.1	101.7	99.5	92.1
1996	92.5	93.6	102.0	91.3	98.0	97.2	98.6	48.6	94.6	99.1	96.9	85.5
1997	88.7	93.3	100.0	97.2	100.9	100.6	103.8	52.7	100.6	106.1	102.9	92.9
1998	94.6	97.1	105.4	98.0	103.4	102.5	105.0	50.9	101.9	104.6	102.6	87.8
1999	93.9	94.8	104.0	95.7	100.2	101.2	104.4	53.1	101.5	106.2	105.2	92.9
2000	95.1	98.8	108.4	99.4	107.5	105.8	107.2	56.8	105.5	107.9	107.8	99.8
2001	98.2	101.4	111.6	98.3	107.2	105.6	106.8	53.9	102.9	105.5	101.4	93.4
2002	93.7	98.6	106.8	96.0	106.3	103.2	107.1	53.1	103.1	105.0	104.0	93.2
2003	94.7	97.6	106.3	97.6	101.6	101.0	107.6	54.2	101.2	105.5	104.3	93.4

serie per i diversi anni (Figura R 2.8 in alto a destra) oppure mediante il corrispondente grafico polare (Figura R 2.8 in basso a destra).

□

• Le *serie territoriali* (o serie spaziali) esprimono la distribuzione di una variabile in rapporto ad unità statistiche caratterizzate da una dimensione territoriale (comuni, distretti, regioni, stati, etc.).

Le serie territoriali possono essere multiple se le variabili rilevate sono più di una. Se, invece, l'analisi è condotta congiuntamente rispetto allo spazio e al tempo si parla di *serie spazio-temporali*, a loro volta univariate o multivariate.

○ Questi studi sono divenuti importanti nelle analisi dei sistemi ecologici e per la valutazione dell'impatto ambientale delle innovazioni prodotte su un territorio circoscritto. Sono altresì utili per valutare la distribuzione territoriale di variabili concernenti problematiche socio-sanitarie, politiche ed economiche, visualizzandone dimensioni e variazioni nello spazio e nel tempo.

○ Sono oggi disponibili software assai elaborati (GIS = *Geographical Information Systems*) che consentono di riprodurre immediatamente sulle mappe del territorio le informazioni statistiche, in modo da creare una immediata corrispondenza tra i dati e la loro rappresentazione geografica.

Esempio 2.15

○ La Figura 2.1 mostra, utilizzando sfumature differenti (nell'originale, colori differenti) l'intensità della variabile «Grado di sismicità al 31/12/2002 dei comuni italiani». Essa visualizza immediatamente la dinamica del fenomeno rispetto al territorio e consente un'analisi dei problemi connessi molto più efficacemente delle informazioni originarie, che potrebbero essere riportate mediante una lunga elencazione.

□

2.7 Considerazioni finali

La disamina precedente dovrebbe aver mostrato che la disponibilità e la struttura delle informazioni statistiche determinano anche le successive analisi metodologiche.

• In una *serie storica* l'interesse prioritario dello studioso è rivolto al meccanismo dinamico che collega i dati osservati sequenzialmente: in altri termini, si cerca «il processo generatore delle osservazioni temporali» perché la conoscenza di tale evoluzione consente di pervenire alla previsione, alla simulazione e al controllo. In tali analisi, la caratteristica fondamentale è l'ordine consecutivo e unidirezionale del tempo, rispetto al quale il fenomeno viene misurato, studiato, interpretato.

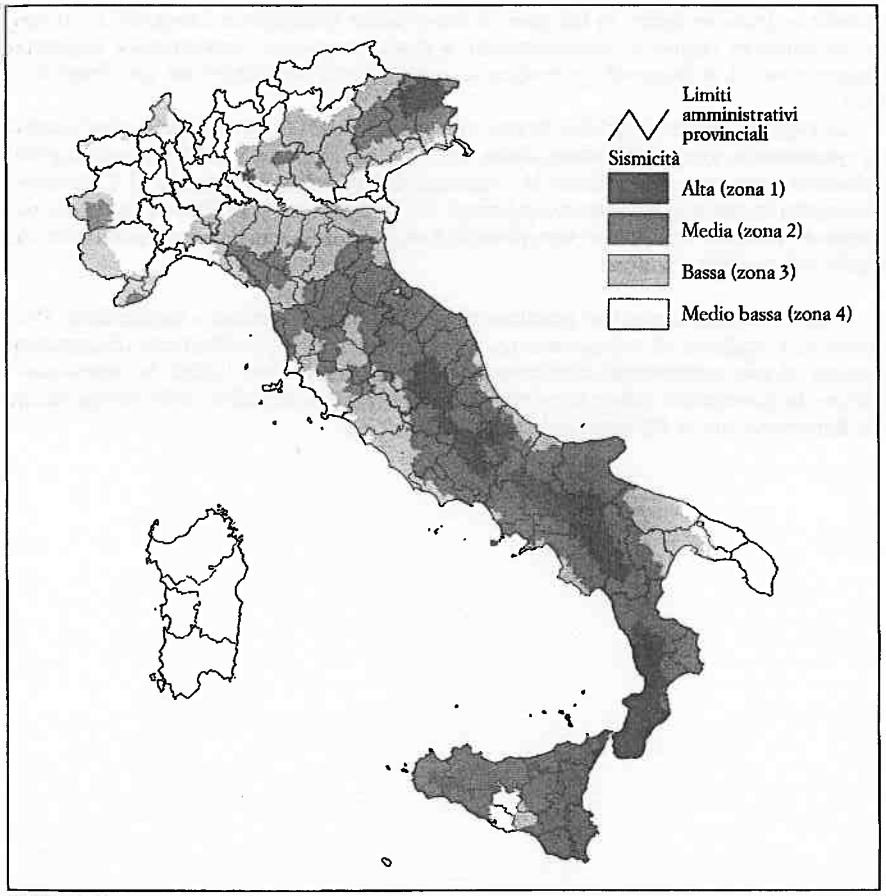

FIGURA 2.1. Comuni d'Italia per grado di sismicità al 31/12/2002 (Annuario Statistico Italiano, ISTAT, 2003).

• Nelle *serie territoriali* viene meno la direzione univoca e il concetto di ordinamento va riferito a quello, meno stringente, di distanza da un centro, in tutte le direzioni. La complessità di tali metodi si accresce perché il territorio è «dato», con mutevolezze e discontinuità intrinseche (mari, fiumi, monti, pianure, etc.) che restringono la «possibilità dei fenomeni di disporsi secondo funzioni-obiettivo».

• Nella *matrice dei dati* le analisi statistiche si dividono in modo abbastanza netto a seconda che l'obiettivo riguardi indagini sulle variabili (cioè le colonne: in tal caso, diviene rilevante la connessione tra di esse, la ridondanza, la capacità informativa e previsiva, la individuazione di legami spuri, etc.) o sulle unità

statistiche (cioè le righe: in tal caso, è importante distinguere i soggetti in gruppi, individuare regole di classificazione e discriminazione, determinare soggetti rappresentativi o anomali, pervenire a spazi di rappresentazione più leggibili, etc.).

La rappresentazione grafica di una matrice dei dati (per evidenziare similitudini, connessioni, gruppi, caratteristiche, etc.) pone complessi problemi matematici e statistici non essendo univoca la riduzione di informazioni con più di 3 variabili/soggetti in spazi di dimensioni inferiori. La letteratura specialistica ha prodotto numerosi metodi, ma la loro complessità è superiore a quella presupposta per lo studio del presente volume.

- Altre rappresentazioni grafiche riguardano dati di natura particolare. Per esempio, i *rapporti di composizione*, per i quali la rappresentazione triangolare agevola alcune interessanti interpretazioni; i *dati ciclici*, per i quali la rappresentazione in coordinate polari consente di visualizzare la stabilità delle componenti del fenomeno che si ripetono nel tempo, e così via.

Ambiente R

Rilevazioni statistiche

R2.1 Gestione delle variabili in R

Rispetto alla terminologia statistica utilizzata nel testo, R distingue la natura delle variabili in due grandi categorie: numeriche e non-numeriche (le quali a loro volta si distinguono in sotto-categorie). Quindi, è importante classificare le variabili a seconda della loro natura, per imparare poi ad organizzarle ed analizzarle. In un primo momento, la natura delle variabili è definita dalle modalità che possono assumere; in un secondo momento, le modalità delle variabili si organizzano in oggetti (*objects*), che sono il nucleo centrale di qualsiasi analisi in R .

Le variabili statistiche quantitative (discrete e continue) sono considerate in R tutte *numeriche* ed è praticamente irrilevante per tale ambiente se assumono valori interi, decimali, irrazionali (od anche complessi): la distinzione riguarda, invece, lo statistico che deve conoscere la correttezza delle analisi che può condurre, come più volte richiamato nel testo.

Alcuni concetti fondamentali riguardano *class* e *mode* degli oggetti di R (spesso, gli errori derivano dal fatto che si applicano funzioni a categorie non applicabili). Tuttavia, tali nozioni più complesse non sono strettamente essenziali per le pagine successive; peraltro, la loro comprensione richiede elementi che introdurremo via via in queste note.

R2.2 Natura delle variabili in R

Con riguardo alla loro natura, le variabili in R si distinguono in:

numeriche Sono variabili le cui modalità sono numeri (età, reddito, consumi, investimenti, velocità, prezzi, voto, lunghezza, durata, umidità, altezza, peso, ...).

stringhe Sono variabili le cui modalità sono alfanumeriche, cioè composte da lettere eventualmente con numeri (città di nascita, regione di residenza, corso di laurea prescelto, diploma secondario, professione, codice fiscale,...). Le modalità delle stringhe si indicano tra " " oppure tra ' '.

fattori Sono le variabili qualitative (nominali ed ordinali), le cui modalità in R si chiamano livelli (*levels*). Tali modalità in effetti sono stringhe (anche se R , al suo interno,

le considera come se fossero numeri) ma con esse si può operare e produrre analisi statistiche (come vedremo). In assenza di un ordine imposto dall'utente, l'ordinamento delle modalità è alfabetico.

logiche Sono fattori che assumono due sole modalità: vero (=TRUE oppure T) e falso (=FALSE oppure F) e sono il risultato della verifica di una affermazione. In alcuni casi, R attribuisce a TRUE e FALSE i valori numerici 1 e 0, rispettivamente.

dato mancante Sono valori indicati con NA (=*not available*) ed esprimono l'assenza di una modalità (*missing value*) per una o più unità statistiche. È possibile includerli od escluderli dalle analisi mediante opportune istruzioni.

operazioni atipiche Sono indicati con NaN (=*not a number*) quelle operazioni matematiche che sono indeterminate (come 0/0 oppure $\infty - \infty$ e ∞/∞). Invece, R riconosce ∞ come un numero e lo indica con `Inf` (anche con segno, se richiesto), per cui sono ammissibili in R (ma non in Matematica!) calcoli come $1/0 =$`Inf` e $\log(0) = -$`Inf`. Il controllo di tali aspetti è talvolta necessario per evitare che elaborazioni automatiche si blocchino per la presenza di un NaN o producano risultati inutilizzabili per la presenza di ∞. Si noti che, secondo R , $0^0 = 1$, mentre 0^0 è una forma indeterminata in Matematica; invece, è corretto scrivere: $\lim_{x \to 0+} x^x = 1$.

In R è possibile assegnare la natura ad una variabile, richiederla, modificarla: nei casi in cui queste operazioni saranno necessarie verrà evidenziato di volta in volta. Qui ci limitiamo a segnalare solo che R è fornito di due gruppi di apposite funzioni:

- Le funzioni che iniziano con `is....` (come: `is.numeric`, `is.factor`, `is.character`, `is.list`, etc.) mediante le quali *si chiede* ad un oggetto se è vero che è del tipo affermato. Si osservino gli esempi:

```
R==> is.numeric(2); is.character("AAA");
[1] TRUE
[1] TRUE
R==> is.numeric("AAA"); is.character(2);
[1] FALSE
[1] FALSE
```

- Le funzioni che iniziano con `as....` (come: `as.numeric`, `as.factor`, `as.character`, `as.list`, etc.) mediante le quali *si costringe* un oggetto ad essere del tipo affermato, modificandolo se non lo è. Si osservino gli esempi:

```
### costringe 25 ad essere una modalità qualitativa
R==> as.character(25);
[1] "25"
### costringe il valore logico FALSE ad essere qualitativo
R==> as.character(3==4);
[1] "FALSE"
### costringe il valore logico TRUE ad essere un numero
R==> as.numeric(2+3==5);
[1] 1
### costringe la modalità qualitativa "AAA" ad un numero
### ma ciò provoca in R un NA, cioè un valore mancante
### ed un avviso (=Warning message)
R==> as.numeric("AAA");
[1] NA
Warning message:
si è prodotto un NA per coercizione
```


In assenza di specifiche indicazioni, R mette in ordine alfabetico le modalità delle variabili qualitative; desiderando un ordinamento differente esso va specificato come nell'esempio seguente:

```
R==> aaa=c("Daniela","Beatrice","Anna","Chiara");
R==> ordered(aaa,levels=c("Anna","Beatrice","Chiara","Daniela"));
[1] Daniela  Beatrice Anna    Chiara
Levels: Anna < Beatrice < Chiara < Daniela
```

R2.3 Gli oggetti di R

L'ambiente R è fondato su un gran numero di *oggetti* ma, nel seguito, ne utilizzeremo solo alcuni: vettori, matrici, *data frame* e, occasionalmente, liste. Altri oggetti (come le date, per esempio) sono importanti in contesti più specifici e non saranno qui affrontati.

Gli oggetti sono i contenitori delle variabili prima definite e, nei loro tratti essenziali, sono così caratterizzati:

vettori Sono un elenco ordinato di modalità (quantitative, cioè numeri, oppure qualitative) e possono essere manipolati come se fossero entità singole. È fondamentale che tutti gli elementi di un vettore siano *dello stesso tipo* (cioè tutti numerici, tutte modalità logiche, tutte modalità qualitative, etc.). La dimensione del vettore è la sua lunghezza, cioè il numero di elementi che lo compongono.

matrici Sono un elenco ordinato di modalità (solo quantitative, solo qualitative, solo logiche) organizzate per righe e per colonne. È fondamentale che tutti gli elementi di una matrice siano *dello stesso tipo*. La dimensione di una matrice è costituita dal numero di righe e di colonne che la compongono.

data frame Sono un elenco di informazioni statistiche costituite tutte dallo stesso numero di righe e colonne (come una matrice); generalmente, ciascuna riga è riferita ad una unità statistica e ciascuna colonna rappresenta una variabile. È ammesso che le colonne di un data frame siano *di natura differente* (cioè taluni numerici, altri qualitativi, altri logici, etc.). La dimensione di un data frame è costituita dal numero di righe e di colonne che lo compongono.

liste Sono un *elenco di oggetti che possono essere di natura differente* e che vengono accomunati perché connessi ad una medesima analisi statistica. Per esempio, nel Capitolo 18, vedremo una lista di oggetti di natura differente ma tutti collegati ad un modello di regressione.

R2.4 Analisi di un vettore

Le variabili univariate discusse in questo Capitolo del testo si analizzano mediante vettori; quindi, pur senza addentrarci nell'algebra lineare, introdurremo le principali operazioni tra vettori in R .

Esistono molti modi di costruire e/o definire vettori e, in assenza di indicazioni, i vettori si intendono come una lista di elementi disposti per colonna. Per avere vettori riga da vettori colonna occorre utilizzare il comando `t()` che produce la trasposta di un vettore; viceversa, applicando la trasposta ad un vettore riga si ottiene un vettore colonna.

Per conoscere la dimensione di un vettore `vett`, cioè la sua lunghezza, si utilizza la funzione `length(vett)`.

Gli esempi ed i commenti seguenti illustrano le differenti istruzioni:

```
R==> v1=1:10; v1            # poni in v1 i numeri interi da 1 a 10,
                            # e poi mostra v1
[1]  1  2  3  4  5  6  7  8  9 10

R==> v2=seq(2,20,2); v2     # poni in v2 gli interi da 2 a 20,
                            # di 2 in 2, e mostra v2
[1]  2  4  6  8 10 12 14 16 18 20

R==> v3=seq(from=2,to=20,by=2); v3 # idem, come sopra, e mostra v3
[1]  2  4  6  8 10 12 14 16 18 20

R==> length(v1);   length(1:100);  # dimensione dei vettori
[1] 10
[1] 100

R==> v4=seq(20,2,-2); v4    # poni in v4 gli interi da 20 a 2,
                            # decrescenti di 2 in 2, e mostra v4
[1] 20 18 16 14 12 10  8  6  4  2
```

Se si desidera costruire l'elenco dei decimali nell'intervallo tra 0 e 1, si ha:

```
R==> yy=(0:10)/10;  yy      # poni in yy (e mostra):
                            # 0.0, 0.1, ..., 1.0
[1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
```

A tale riguardo va sottolineato come il simbolo : (che genera una sequenza ordinata) abbia la precedenza sulle altre operazioni, per cui bisogna esercitare una particolare attenzione nel suo uso combinato. Si considerino, a conferma, i seguenti esempi numerici:

```
R==> 1:10-1;     # equivale a (1:10)-1, cioè sottrarre 1
                 # a ciascuno dei numeri da 1 a 10
[1] 0 1 2 3 4 5 6 7 8 9

R==> 1:(10-1)  # equivale a 1:9
[1] 1 2 3 4 5 6 7 8 9
```

Un'ulteriore possibilità di definire un vettore (nel quale uno o gruppi di valori si ripetono) è tramite la funzione rep il cui uso è esemplificato nelle seguenti istruzioni:

```
R==> rep(1,5);      # replica 1 per 5 volte
[1] 1 1 1 1 1
R==> rep(1:3, 4)    # replica 1:3 per 4 volte
[1] 1 2 3 1 2 3 1 2 3 1 2 3
R==> rep(1:5, each = 2)  # replica ciascun elemento
                         # di 1:5 per 2 volte
[1] 1 1 2 2 3 3 4 4 5 5
R==> rep(1:4, c(1,2,3,4)) # replica come nel secondo vettore
[1] 1 2 2 3 3 3 4 4 4 4
R==> rep(1:3, each = 2, times = 3) # replica 3 volte la ripetizione
 [1] 1 1 2 2 3 3 1 1 2 2 3 3 1 1 2 2 3 3
```

In R sono anche definite le 26 lettere dell'alfabeto (Inglese) che sono una sequenza ordinata di caratteri. L'esempio mostra come generarne alcune, rispettivamente, in minuscolo o maiuscolo:

```
R==> letters[1:13];
```



```
 [1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j" "k" "l" "m"
R==> LETTERS[14:26];
 [1] "N" "O" "P" "Q" "R" "S" "T" "U" "V" "W" "X" "Y" "Z"
```

Se gli elementi di un vettore sono pochi, è semplice concatenarli mediante l'operatore c() nel modo seguente:

```
R==> var1=c(1,2,4,7,-1,-12,0,2,5); var1
[1]   1   2   4   7  -1 -12   0   2   5
```

L'operatore c() può essere applicato anche a vettori già definiti ed è ricorsivo, cioè può essere applicato più volte, in modo concatenato. Per esempio:

```
R==> var2=c(0,0,-1,var1,8); var2
[1]   0   0  -1   1   2   4   7  -1 -12   0   2   5   8
R==> var3=c(var1,c(1,2,3)); var3
[1]   1   2   4   7  -1 -12   0   2   5   1   2   3
```

L'accesso ai singoli elementi di un vettore già definito avviene mediante una indicizzazione che si effettua con le parentesi quadre, e la cui versatilità può essere apprezzata con gli esempi:

```
R==> var1[3]
[1] 4
R==> var1[1:3]
[1] 1 2 4
R==> var1[c(1,3,4,7,2,1)]
[1] 1 4 7 0 2 1
R==> var1[9:1]
[1] 5   2   0 -12  -1   7   4   2   1
R==> var1[seq(2,9,by=3)]
[1]  2 -1  2
```

È anche semplice selezionare gli elementi di un vettore che obbediscono ad una condizione nel modo seguente:

```
R==> var1[var1>0]   # seleziona i valori di var1 maggiori di 0
[1] 1 2 4 7 2 5
R==> var1[var1<=0]  # seleziona i valori di var1 non-positivi
[1]  -1 -12   0
R==> var1[var1==2]  # seleziona i valori uguali a 2
[1] 2 2
R==> which(var1==2)  # seleziona gli indici dei valori uguali a 2
[1] 2 8
```

L'ultima funzione `which(condizione)` elenca gli indici del vettore che soddisfano una `condizione` (logica), ed ha molte applicazioni nell'analisi dei dati statistici.

Infine, si possono anche escludere elementi da un vettore anteponendo l'operatore `-()` agli indici da escludere:

```
R==> var2
[1]   0   0  -1   1   2   4   7  -1 -12   0   2   5   8
R==> var2[-c(1,2,4,7)]  # escludi var2[1],var2[2],var2[4],var2[7]
[1]  -1   2   4  -1 -12   0   2   5   8
```

Una caratteristica di R (e di molti linguaggi simili) è la possibilità di operare su vettori come se fossero quantità algebriche tutte le volte che le operazioni richieste sono eseguite elemento-per-elemento (*elementwise*). Quindi, 3*vett moltiplicherà per 3 tutti gli

elementi di vett; invece, 2.5+vett aggiungerà il numero 2.5 a tutti gli elementi di vett. Questa caratteristica – che sul piano formale dell'algebra lineare non è sempre corretta – si chiama *recycling* (perché R ricicla gli elementi di cui dispone sino ad operare con vettori di pari lunghezza) ed è di per sé molto utile.

Quindi, in R sono lecite le istruzioni seguenti:

```
R==> x=c(1,2,3,4,1); y=c(10,8,9,12,7);
R==> x+y;
[1] 11 10 12 16  8
R==> y-x;
[1] 9 6 6 8 6
R==> x*y;
[1] 10 16 27 48  7
R==> y/x;
[1] 10  4  3  3  7
R==> y^x
[1]    10    64   729 20736     7
```

Purtroppo, ciò genera anche qualche problema perché si può ottenere un risultato (che sembra apparentemente valido) anche se si mettono assieme operazioni scorrette sul piano sostanziale (per esempio, se si sommano vettori di lunghezza differente); talvolta, R aggiunge al risultato un *warning* (=avvertimento), come nell'esempio seguente:

```
R==> xx=1:5; yy=c(10,11,12);  xx+yy
[1] 11 13 15 14 16
Warning message:
In xx + yy : longer object length is not
a multiple of shorter object length
```

Dati due vettori *di pari lunghezza* sono lecite in R le operazioni algebriche binarie:
vett1+vett2, vett1-vett2, vett1*vett2,vett1/vett2
intese come operazioni tra elementi corrispondenti.

Infine, gli elementi di un vettore si ordinano in senso non-decrescente con la funzione sort(); se si specifica l'opzione decreasing=TRUE l'ordinamento diventa non-crescente, come esemplificato con le seguenti istruzioni:

```
R==> vettore=c(19,17,11,2,5,1,8,10,4,3,7,14,15,12,6,20,9,18,16,13);
### ordina in senso non-decrescente gli elementi di vettore
R==> sort(vettore);
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
### ordina in senso non-crescente  gli elementi di vettore
R==> sort(vettore,decreasing=TRUE);
 [1] 20 19 18 17 16 15 14 13 12 11 10  9  8  7  6  5  4  3  2  1
```

Il contenuto di un vettore può anche consistere in variabili logiche (cioè TRUE e FALSE) come nell'esempio seguente:

```
R==> vettlogico=c(2==2,3==4,4<3,4>7,10==5*2,2<5);  vettlogico;
[1]  TRUE FALSE FALSE FALSE  TRUE  TRUE
```

Le operazioni logiche di "e" (nel senso di "sia...sia" ed indicata con il simbolo &) e "o" (nel senso di "o...o"ed indicata con il simbolo |) si utilizzano tra elementi corrispondenti dei vettori come nelle istruzioni seguenti:

```
R==> vett1=c(TRUE,TRUE,FALSE); vett2=c(FALSE,TRUE,TRUE);
R==> vett1&vett2;
[1] FALSE  TRUE FALSE
R==> vett1|vett2;
[1] TRUE TRUE TRUE
```


Infatti, TRUE "e" FALSE genera FALSE, mentre TRUE "o" FALSE genera TRUE.

R2.5 Analisi di una variabile quantitativa

Una variabile quantitativa le cui osservazioni (modalità) sono raccolte consecutivamente si chiama *serie*. Quando le informazioni sono organizzate in una tabella trattasi di una *seriazione* o *distribuzione di frequenza* (vedi paragrafo 2.2 del testo).

Le modalità di una variabile quantitativa, disponibile in serie, possono essere inserite direttamente dalla tastiera quando la loro numerosità è limitata tramite la funzione c(). Così, la variabile altezza di 10 studenti si crea mediante:

```
R==> altezza=c(180,180,173,187,164,186,175,170,180,170);
R==> min(altezza); max(altezza); which(altezza==173);
[1] 164
[1] 187
[1] 3
```

Le funzioni min(), max(), which() forniscono, rispettivamente, l'altezza minima, l'altezza massima e l'indice dello studente al quale corrisponde un'altezza pari a 173 cm: in questo esempio, è il terzo.

Se i dati sono immessi da tastiera, si può attivare la funzione scan() e poi digitare le osservazioni una dopo l'altra inviando Enter dopo ogni dato, e inviando solo Enter per concludere l'immissione dati. Così, la sequenza di istruzioni:

```
R==> altezza=scan()   # poi il tasto Enter
R==> 1: 180   # poi il tasto Enter (il numero 1: è fornito da R)
R==> 2: 180   # poi il tasto Enter (il numero 2: è fornito da R)
R==> 3: 173   # poi il tasto Enter (il numero 3: è fornito da R)
R==> 4: 187   # poi il tasto Enter
R==> 5: 164   # poi il tasto Enter
R==> 6: 186   # poi il tasto Enter
R==> 7: 175   # poi il tasto Enter
R==> 8: 170   # poi il tasto Enter
R==> 9: 180   # poi il tasto Enter
R==> 10: 170  # poi il tasto Enter
R==> 11:      # utilizzare il tasto Enter per concludere
Read 10 items
```

costruisce il medesimo vettore altezza prima inserito tramite c().

Se, invece, i dati sono conservati in un file (per esempio, il file: prezzolibri.txt contiene la variabile prezzo nel data set prezzolibri), l'istruzione da utilizzare è la seguente:

```
R==> dati=read.table("prezzolibri.txt", header=TRUE);
```

In tal modo, i dati contenuti nel file verranno immessi nel vettore prezzo e sono utilizzabili per le analisi statistiche. L'aggiunta header=TRUE indica che la prima riga del file contiene il nome della variabile. Utilizzando la funzione attach() le singole colonne di dati sono utilizzabili come vettori con il loro nome corrispondente. Così, sono lecite le istruzioni:

```
R==> attach(dati);
R==> length(prezzo);
[1] 420
R==> min(prezzo); max(prezzo);
[1] 12
```

```
[1] 100
R==> length(which(prezzo>40)); # quante modalità hanno (prezzo>40)
[1] 151
```

Esse, nell'ordine, mostrano che le modalità contenute nel vettore (si tratta di osservazioni in serie) sono in numero di 420, il valore minimo è 12 e quello massimo è 100 Euro; infine, sono 151 i libri che hanno un prezzo superiore a 40 Euro.

Si suggerisce di utilizzare *sempre* `detach(dati)` quando le analisi concernenti le variabili contenute in `dati` sono concluse. Infatti, anche se `attach(dati)` consente di utilizzare le variabili con lo stesso nome con il quale sono definite nel file di `dati`, esso crea conflitti quando si operano trasformazioni. Esistono alternative meno problematiche, come le seguenti istruzioni, che producono gli stessi risultati:

```
R==> with(data=dati,length(prezzo));
[1] 420
R==> with(data=dati,min(prezzo));with(data=dati,max(prezzo));
[1] 12
[1] 100
R==> with(data=dati,length(which(prezzo>40)));
[1] 151
```

È opportuno visualizzare le prime righe del file letto tramite la funzione `head()`, come qui esemplificato per il data set `prezzolibri`:

```
R==> dati=read.table("prezzolibri.txt", header=TRUE);
R==> head(dati);
  prezzo
1     24
2     20
3     28
4     52
5     58
6     34
```

Similmente, la funzione `tail()` fornisce le ultime righe del vettore. Invece, la funzione `str()` evidenzia la struttura dei dati e la caratteristica delle variabili, fornendo anche i primi valori:

```
R==> dati=read.table("prezzolibri.txt", header=TRUE);
R==> str(dati);
'data.frame':   420 obs. of  1 variable:
 $ prezzo: int  24 20 28 52 58 34 18 16 16 18 ...
```

In tal caso, si evidenzia che l'oggetto `dati` contiene il vettore `prezzo` consistente in 420 osservazioni che `R` riconosce come modalità intere. Se si vuole accedere alla variabile `prezzo`, occorre digitare `dati$prezzo`; in tal modo, si ottiene una visualizzazione della variabile nella finestra di `R` :

```
R==> dati$prezzo
  [1]  24  20  28  52  58  34  18  16  16  18  18  45  70  20  65
20  18  34
 [19]  28  18  18  18  28  25  22  60  34  32  44  24  35  36  65
24  48  48
 [37]  .....................................
[397]  35  18  50  30  28  54  22  50  28  54  35  35  16  40  40
38  18  52
[415]  80  30  30  36  42  32
```


Si noti come, essendo un vettore, `R` visualizza le modalità di seguito, indicando all'inizio di ciascuna riga l'indice della prima unità statistica di riferimento; quindi, 24 è il prezzo del primo libro, 28 è il prezzo del diciannovesimo, ..., 32 è il prezzo dell'ultimo libro, cioè il 420° nella lista.

R2.6 Matrici

Le matrici sono una collezione di *elementi della stessa natura* organizzati per righe e per colonne. La funzione che le definisce è `matrix` il cui utilizzo è esemplificato qui di seguito:

```
R==> matrix(3,nrow=2,ncol=3)
     [,1] [,2] [,3]
[1,]    3    3    3
[2,]    3    3    3

R==> matrix(1:6,nrow=2,ncol=3)
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6

R==> matrix(1:6,nrow=2,ncol=3,byrow=TRUE)
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6

R==> xx=c(1:6,-3,0,1,9:7); matrix(xx,nrow=3,ncol=4)
     [,1] [,2] [,3] [,4]
[1,]    1    4   -3    9
[2,]    2    5    0    8
[3,]    3    6    1    7

R==> matrix(xx,nrow=3)
     [,1] [,2] [,3] [,4]
[1,]    1    4   -3    9
[2,]    2    5    0    8
[3,]    3    6    1    7
```

La prima matrice è fatta di costanti tutte pari a 3; la seconda colloca la sequenza `1:6` *secondo le colonne*. Per creare un ordinamento secondo le righe (come è usuale in algebra ed in altri linguaggi) occorre specificare l'istruzione `byrow=TRUE`. La penultima istruzione mostra come comporre in `R` gli elementi di una matrice `xx` mediante `c()`. L'ultima istruzione mostra che non è strettamente necessario specificare sia il numero di righe che di colonne di una matrice, perché la dimensione non specificata viene determinata dalla lunghezza del primo elemento della funzione `matrix` (ma, se si sbaglia per omissione, `R` ricicla i dati utilizzando, alla fine, i primi elementi). Quindi, per esempio, addizionando una costante ad una matrice, `R` crea una matrice di costanti pari alla prima, e così via. In tutti i casi, si noti come `R` indicizza gli elementi di una matrice, specificando la riga e la colonna di riferimento.

Le matrici possono anche essere costruite affiancando vettori di uguale lunghezza. A tal fine, le funzioni `cbind()` e `rbind()` uniscono vettori colonna (con lo stesso numero di righe) e vettori riga (con lo stesso numero di colonne), rispettivamente, come ora esemplifichiamo:

```
R==> a=1:4; b=2*a-5; d=a^2; effe=cbind(a,b,d); effe
```

```
     a  b  d
[1,] 1 -3  1
[2,] 2 -1  4
[3,] 3  1  9
[4,] 4  3 16

R==> at=t(a); bt=t(b); dt=t(d); effet=rbind(at,bt,dt); effet
     [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]   -3   -1    1    3
[3,]    1    4    9   16
```

Si osservi che `nrow=` e `ncol=` sono argomenti della funzione `matrix()` per specificarne numero di righe e di colonne; ma `nrow()` e `ncol()` sono anche funzioni che producono in output il numero di righe e di colonne della matrice cui sono applicate. Si veda l'esempio:

```
R==> yyy=matrix(1:12,nrow=3,ncol=4); yyy;
     [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
R==> nrow(yyy);
[1] 3
R==> ncol(yyy);
[1] 4
```

La funzione `diag` ha effetti diversi a seconda dell'argomento cui viene applicata:

```
R==> diag(1,nrow=4,ncol=4)
     [,1] [,2] [,3] [,4]
[1,]    1    0    0    0
[2,]    0    1    0    0
[3,]    0    0    1    0
[4,]    0    0    0    1

R==> diag(c(-1,-2,0,5),nrow=4,ncol=4)
     [,1] [,2] [,3] [,4]
[1,]   -1    0    0    0
[2,]    0   -2    0    0
[3,]    0    0    0    0
[4,]    0    0    0    5

R==> bb=matrix(1:16,4,4); diag(bb)
[1]  1  6 11 16
```

Come si vede, nel primo caso, `diag` ha creato una matrice 4×4 collocando 1 nella diagonale principale; nel secondo caso ha posto il vettore dato nella diagonale principale; nel terzo caso ha estratto dalla matrice `bb` la diagonale principale.

Per creare una (piccola) matrice di dati tramite immissione manuale da tastiera si può utilizzare la funzione `scan()` (già incontrata nel paragrafo precedente) all'interno di `matrix()`. Esemplifichiamo il caso di quattro informazioni numeriche per riga:

```
R==> matrice=matrix(scan(),ncol=4, byrow=TRUE);
1: 1 3 2 90        ### digitare 1 3 2 e 90, poi Enter
5: 2 3 4 80        ### digitare 2 3 4 e 80, poi Enter
9: 3 2 1 70        ### digitare 3 2 1 e 70, poi Enter
13:                ### immettere Enter per concludere
```



```
Read 12 items      ### R informa che ha immesso 12 elementi
R==> matrice;
     [,1] [,2] [,3] [,4]
[1,]    1    3    2   90
[2,]    2    3    4   80
[3,]    3    2    1   70
```

Nel paragrafo seguente mostreremo modi più efficienti per immettere dati, quando sono numerosi (come nei data set) e non si intende correre il rischio di commettere errori di digitazione.

Altri comandi e funzioni su matrici saranno utilizzati all'occorrenza; ripetiamo che `t(X)` effettua la trasposizione di righe e colonne di qualsiasi matrice X; invece, `det(Y)` e `solve(Y)` effettuano, rispettivamente, il determinante e l'inversa di una matrice quadrata Y.

Va ripetuto che le operazioni algebriche in `R` avvengono tra elementi corrispondenti dei vettori e delle matrici (quando sono lecite), anche se talvolta sono matematicamente scorrette. Invece, il prodotto matriciale standard (cioè righe per colonne tra vettori/matrici compatibili) si effettua tramite l'operazione indicata con `%*%`. Nei casi compatibili, ciò vale anche per il prodotto riga × colonna e colonna × riga, come negli esempi seguenti:

```
R==> x=c(1,2,3); y=c(10,8,9);
R==> t(x)%*%y; ### il prodotto x'y è uno scalare
     [,1]
[1,]   53
R==> x%*%t(y); ### il prodotto x*y' è una matrice
     [,1] [,2] [,3]
[1,]   10    8    9
[2,]   20   16   18
[3,]   30   24   27
```

Infine, si osservi che per sommare le singole righe/colonne di una matrice non si può utilizzare la funzione `sum` (che fornisce la somma di tutti gli elementi della matrice) ma apposite funzioni, come appresso esemplificato:

```
R==> xx=matrix(1:6,nrow=2); xx;
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
R==> sum(xx);          ### somma di tutti gli elementi di xx
[1] 21
R==> colSums(xx);      ### somma degli elementi di ogni colonna
[1]  3  7 11
R==> rowSums(xx);      ### somma degli elementi di ogni riga
[1]  9 12
```

In queste operazioni è possibile escludere la eventuale presenza di dati mancanti (inserendo l'opzione `na.rm=TRUE`).

R2.7 *Data frame*

Secondo la terminologia statistica, la *matrice di dati* è una matrice dal punto di vista della struttura le cui variabili, però (a differenza delle matrici numeriche), possono essere anche di natura diversa. Per questo, viene definito l'oggetto *data frame*, molto utile nelle analisi statistiche dove le informazioni sono predisposte in modo che ciascuna riga faccia

riferimento all'*unità statistica* mentre le colonne sono *variabili* (di natura anche diversa tra loro: qualitativa, quantitativa, dicotomiche, logiche, etc.).

Un *data frame* si può creare a partire da variabili (cioè vettori) già esistenti, per esempio mediante le istruzioni:

```
R==> a=c(1,2,3,4,5); b=6:10; c=rep(2,5);
R==> ddd=data.frame(a,b,c); ddd;
  a  b c
1 1  6 2
2 2  7 2
3 3  8 2
4 4  9 2
5 5 10 2
R==> names(ddd)
[1] "a" "b" "c"
```

Come si vede, in assenza di altre dichiarazioni, vengono attribuiti alle colonne del *data frame* i nomi dei vettori componenti ed alle righe i numeri interi. Ma possono essere specificati per il *data frame* ddd sia i nomi delle colonne (=variabili) che delle righe (=unità statistiche) secondo le proprie esigenze:

```
R==> colnames(ddd)=c("X1","X2","X3");
R==> rownames(ddd)=c("Anna","Biagio","Carla","Dario","Emilia");
R==> ddd;
       X1 X2 X3
Anna    1  6  2
Biagio  2  7  2
Carla   3  8  2
Dario   4  9  2
Emilia  5 10  2
```

È possibile operare su sottoinsiemi di righe (mediante una opzione logica del tipo `subset=`) e/o di colonne (mediante una elencazione del tipo `select=...`) utilizzando la funzione `subset()`, nel modo seguente:

```
R==> subset(ddd,subset=(X1<=3));### righe tali che X1 non supera 3
       X1 X2 X3
Anna    1  6  2
Biagio  2  7  2
Carla   3  8  2
R==> subset(ddd,select=c(1,3));    ### solo colonne 1 e 3
       X1 X3
Anna    1  2
Biagio  2  2
Carla   3  2
Dario   4  2
Emilia  5  2
R==> subset(ddd,subset=(X2>8),select=2);### seleziona righe
                                        ### e colonne
       X2
Dario   9
Emilia 10
```

Molto più spesso, la matrice dei dati proviene da un file esterno, già predisposto mediante un editore di testo (file ASCII oppure con estensione *txt*) o fogli elettronici (tipo *Excel*).

■ Se il file da cui leggere è di tipo *testo*, la lettura della matrice dei dati avviene mediante il comando `read.table("nomefile")` nel quale va immessa l'opzione `header=TRUE` se la prima riga contiene i nomi delle variabili: in tal caso, si possono utilizzare isolatamente i vettori da richiamare con lo stesso nome della variabile. Per esempio, le istruzioni seguenti:

```
R==> ### Leggi STUDENTI
R==> dati=read.table("studenti.txt", header=TRUE);
R==> head(dati)
  anno genere  eta peso altezza dipsci dipcla diptec dipalt
1 1988      0 20.6   65     180      0      0      1      0
2 1988      0 20.2   75     180      1      0      0      0
3 1988      0 20.3   60     173      0      0      1      0
4 1988      0 23.9   93     187      1      0      0      0
5 1988      0 21.4   66     164      0      0      1      0
6 1988      0 25.0   84     186      0      0      1      0

  componenti occhiali  fumo   diploma
1          6        0     1       tec
2          4        0     0       sci
3          4        1     0       tec
4          8        0     1       sci
5          5        0     0       tec
6          4        0     0       tec
```

consentono di leggere le informazioni dal file `studenti.txt` e creare le variabili (cioè i vettori) con i nomi `anno, genere, eta, peso, altezza, dipsci, dipcla, diptec, dipalt, componenti, occhiali, fumo, diploma`. Se si intende utilizzare queste variabili (=vettori) con il loro nome occorre utilizzare la funzione `attach(dati)`.

Per indicare con nomi specifici le variabili (=colonne) del *data frame* si può usare l'opzione `col.names=c("nome1","nome2",...)`; in modo analogo, con le righe mediante l'opzione `row.names=c("nome1","nome2",...)`.

In alternativa, alla funzione `attach()` (che, talvolta, può creare conflitti sui nomi delle variabili), si possono definire direttamente le variabili presenti nel *data frame*, nel modo seguente:

```
R==> anno=dati[,1]; genere=dati[,2]; eta=dati[,3]
```

oppure

```
R==> dati$anno; dati$genere; dati$eta
```

Quindi, in generale, una variabile `var` presente nel *data frame* `df` è univocamente specificata mediante `df$df`. Nel seguito, per richiamare le variabili di un *data frame*, utilizzeremo anche la funzione `with()`.

La funzione `dim(dati)` fornisce il numero di righe e colonne del *data frame* `dati`:

```
R==> dim(dati)
[1] 2759   12
```

Nello specifico, essa ci dice che nel data set `studenti` vi sono 2759 osservazioni (=*unità statistiche*, cioè le righe) sulle quali sono state registrate 12 informazioni (=*variabili*, cioè le colonne).

Per individuare gli elementi (singoli, in gruppi, che obbediscono a regole) di un *data frame* ci si avvale delle parentesi quadre: `[ ]` . Quindi, se `dati` è il nome del *data frame*, allora:

- `dati[,2]` indica tutte le modalità della colonna 2;

- `dati[3,]` indica tutte le modalità della riga 3;
- `dati[5,7]` indica il solo elemento collocato alla riga 5 e alla colonna 7;
- `dati[,2:4]` indica tutte le modalità delle colonne 2, 3, 4;
- `dati[2:5,]` indica tutte le modalità delle righe 2, 3, 4, 5;
- `dati[,-c(1:3)]` indica tutte le modalità del *data frame* ad esclusione di quelli delle colonne 1, 2, 3;
- `dati[-c(1:100),]` indica tutte le modalità del *data frame* ad esclusione delle prime 100 righe;
- `dati[1:100,2:4]` indica tutte le modalità del *data frame* che sono presenti nelle prime 100 righe alle colonne 2, 3, 4;
- `dati[,listavar]` indica le modalità del *data frame* di tutte le righe e delle sole colonne elencate in `listavar`;
- `dati[,condizione]` indica tutte le modalità del *data frame* delle sole colonne che obbediscono a `condizione`;

La matrice dei dati contenuta in un *data frame* si può ordinare rispetto ad una o più colonne (cioè variabili) mediante la funzione `order()`. Se il *data frame* si chiama `dati`, allora:

```
### ordina le righe di dati rispetto alla colonna 3
R==> dati[order(dati[,3]),];
### ordina le righe di dati prima rispetto alla colonna 3,
### e poi rispetto alla colonna 5
R==> dati[order(dati[,3],dati[,5]),];
```

In effetti, la funzione `order()` produce una lista di interi secondo l'ordinamento non-decrescente dell'argomento e le istruzioni precedenti applicano queste liste alle righe del *data frame*.

Una situazione utile per creare un *data frame* contenente la combinazione di due vettori numerici (cioè il prodotto cartesiano tra i due insiemi) è offerta dalla funzione `expand.grid` che ora esemplifichiamo:

```
R==> expand.grid(primo=1:3,secondo=2:4);
  primo secondo
1     1       2
2     2       2
3     3       2
4     1       3
5     2       3
6     3       3
7     1       4
8     2       4
9     3       4
```

Quindi, tale funzione genera tutte le coppie ordinate del primo e secondo vettore (facendo variare prima il primo vettore a parità del secondo); essa è molto utile per costruire funzioni di due variabili su una griglia di valori predefiniti.

■ Se il file è di tipo *Excel* esistono comandi diretti per l'immissione e la lettura in `R` ; per evitare di memorizzare molte istruzioni, suggeriamo una procedura molto semplice:

1. Si apre uno *script* nuovo in `R`
2. Si evidenziano i dati in Excel con le frecce della tastiera e si copiano con `Ctrl+C`
3. Si immettono nello *script* portandosi con il cursore in esso e facendo `Ctrl+V`

Questa operazione ha successo se si pone attenzione ad alcuni aspetti:

- Nella prima riga del file *script* si inseriranno i nomi desiderati per le variabili (tanti quanti sono le colonne: non immettere spazi bianchi tra i nomi. Per esempio: "data di nascita" diventa "data.di.nascita"). Non è necessario numerare le righe.
- In Italia, il punto decimale (.) è sostituito dalla virgola (,). Dopo la lettura in `R` del file, occorre trasformare (è immediato con il mouse e tramite il menu) tutte le virgole in punti.
- È buona regola, all'inizio del file, inserire dettagliate descrizioni (precedute da # e che quindi non saranno analizzate) su dati, fonte, variabili, caratteristiche, commenti, etc.

Se la lettura dei dati avviene da un file esistente sul PC ma di cui non si ricordano *directory* e/o nome completo (oppure si teme una errata digitazione) è preferibile utilizzare l'istruzione seguente:

```
R==> dati=read.table(file.choose(),header=TRUE)
```

Essa apre una finestra in Windows (a partire dalla *directory* di lavoro) e consente di individuare il file da cui leggere i dati con una serie di clic, senza preoccuparsi di digitare percorso e nome.

R2.8 Operazioni di input e output in R

Chi lavora con analisi statistiche ha spesso necessità di salvare oggetti e poi di reimportarli per lavorare di nuovo. Occorre distinguere se essi sono oggetti di `R` (e le funzioni da utilizzare sono `save` e `load`) oppure data set prodotti da software più generali (e le funzioni sono `write` and `read`).

■ Per salvare oggetti prodotti da `R` (persino l'intera sessione di lavoro) nel formato leggibile da `R` è sufficiente l'istruzione qui esemplificata, mediante la quale si salva il vettore x, la matrice A, il *data frame* `dafr` e la lista indicata con `list`:

```
R==> save(x,A,dafr,list(aa,bb,cc), file="conserva.rda");
```

L'estensione del file da salvare è `.rda` (ma non è obbligatorio indicarla).

Quando, in una successiva sessione di lavoro, si desidera riportare in memoria tali oggetti occorre l'istruzione:

```
R==> load("conserva.rda");
```

Essa rende disponibili all'ambiente `R` tutti gli oggetti prima salvati con `save()`.

■ Per salvare matrici prodotte in `R` in formato ASCII (leggibili quindi da qualsiasi editore di testo) si utilizza la funzione `write()`, specificando il numero di colonne e facendo la trasposta della matrice originaria per poterla evidenziare per righe (come è standard). Si consideri la seguente esemplificazione:

```
R==> x=matrix(1:18,ncol=6); x;
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    1    4    7   10   13   16
[2,]    2    5    8   11   14   17
[3,]    3    6    9   12   15   18
R==> write(x,ncol=6,file="provascrittura.txt");
```

Se si apre il file `provascrittura.txt` si vede che esso è così composto:

```
1 2 3 4 5 6
7 8 9 10 11 12
13 14 15 16 17 18
```

Se, invece, si utilizzano le istruzioni:

```
R==> write(t(x),ncol=6,file="provascrittura.txt");
```

il file "provascrittura.txt" è correttamente composto come la matrice x:

```
1 4 7 10 13 16
2 5 8 11 14 17
3 6 9 12 15 18
```

L'estensione `.txt` (qui utilizzata) non è obbligatoria ma è comune per i file testo.

Se, invece, si intende salvare in ASCI un intero *data frame* prodotto in R la funzione è `write.table()` (che è l'equivalente in quanto output di quello che è `read.table()` per leggere i *data frame* in input), per la quale si suggerisce di leggere l'aiuto di R per le opzioni. Qui esemplifichiamo una situazione molto semplice:

```
R==> dfx=data.frame(X1=11:15,X2=LETTERS[1:5],
        X3=seq(10,30,length=5));
R==> dfx;
  X1 X2 X3
1  11  A 10
2  12  B 15
3  13  C 20
4  14  D 25
5  15  E 30

R==> write.table(dfx,"provadataframe.txt",quote=FALSE);
```

Se apriamo il file `provadataframe.txt` con un editore di testo (per esempio, in R facendo clic sul menu "File" e poi su "Apri script"), leggiamo:

```
X1 X2 X3
1 11 A 10
2 12 B 15
3 13 C 20
4 14 D 25
5 15 E 30
```

Si osservi che l'opzione quote=FALSE ha evitato di inserire "" ad ogni modalità qualitativa nel file; inoltre, R ha conservato i nomi delle variabili presenti nel data frame (ma si possono cambiare) ed ha ripetuto anche la prima colonna con la sequenza degli interi, usata per numerare le righe (ma si possono dare anche nomi differenti, oppure eliminarli con l'opzione: `row.names=FALSE`).

■ Esistono altri formati di lettura in R (per esempio, file con informazioni delimitate da virgole), ed è anche possibile leggere in R dati provenienti dai più diffusi software commerciali (utilizzando il *package* `foreign`, per esempio); in particolare, è possibile leggere direttamente file Excel senza modificarli previamente.

Segnaliamo, infine, che R consente la lettura di file presenti in Internet in modo efficiente (anche zippati), e per questi approfondimenti rinviamo ai manuali disponibili in rete nonché alla bibliografia del paragrafo 1.7.

R2.9 Rappresentazioni grafiche

L'ambiente R è ricchissimo di strumenti grafici (semplici, raffinati e complessi), per cui quelli qui di seguito introdotti sono i più immediati, quasi tutti realizzabili senza l'attivazione di ulteriori *packages*. Peraltro, sono molteplici i *packages* esplicitamente costruiti per ottenere rappresentazioni grafiche accurate ed efficaci: per chi intende sviluppare l'analisi grafica di dati statistici si consiglia di esaminare i *packages* `grid` e `lattice` e, per i più esigenti, `ggplot2`. Al momento, sono disponibili molti libri (e siti) interamente dedicati alla grafica tramite R .

Concetti generali

Una distinzione fondamentale in R è tra funzioni di alto livello (cioè primarie, *high-level plots*) e di basso livello (cioè secondarie, *low-level plots*). Brevemente, le prime sono essenziali per definire l'ambiente grafico in cui si desidera lavorare e sono autonome; invece, le seconde operano solo se già esiste un grafico generato con le funzioni *high-level*.

Operativamente, le funzioni grafiche secondarie raffinano, completano o specificano grafici già definiti mediante le funzioni grafiche primarie. Per un elenco, citiamo le più diffuse:

- funzioni primarie: `plot, barplot, pie, hist, boxplot, ...`
- funzioni secondarie: `points, lines, text, abline, polygon, legend, identify, locator, box, ...`

Esse saranno via via esaminate in funzione della loro successiva utilizzazione, per cui iniziamo la presentazione mediante un grafico elementare (Figura R 2.1) nel quale si rappresenta la sequenza di 15 osservazioni numeriche, per esempio l'età in anni compiuti delle persone intervistate in una certa indagine.

```
R==> eta=c(18,19,26,39,60,31,38,50,18,57,68,30,44,67,30);
R==> plot(eta)
```

Come si vede, in assenza di ogni opzione, R si limita a rappresentare le modalità numeriche della variabile `eta` con dei punti isolati e considera come ordinate tali valori; inoltre, appone sugli assi la dizione `Index` per le ascisse ed il nome della variabile `eta` per le ordinate; nessun titolo è specificato per tale grafico. Per contro, nel seguito, vedremo come specificare in ogni dettaglio un grafico per renderlo utile alla sua specifica finalità di rappresentazione di informazioni statistiche.

Buona parte degli elementi caratterizzanti di un grafico sono contenuti nella funzione `par()` le cui molteplici opzioni possono essere consultate mediante `?par`. Per questo, vale un principio generale: la modifica di un parametro grafico resta attiva sino alla introduzione di una nuova modifica, per cui la strategia ottimale è di fare in modo di ritornare ai parametri standard di R dopo aver utilizzato opportune varianti.

Come esempio, modifichiamo sia il colore (in genere, è `col="black"`) che l'ampiezza delle linee (in genere, è `lwd=1`); allora, le istruzioni sono le seguenti:

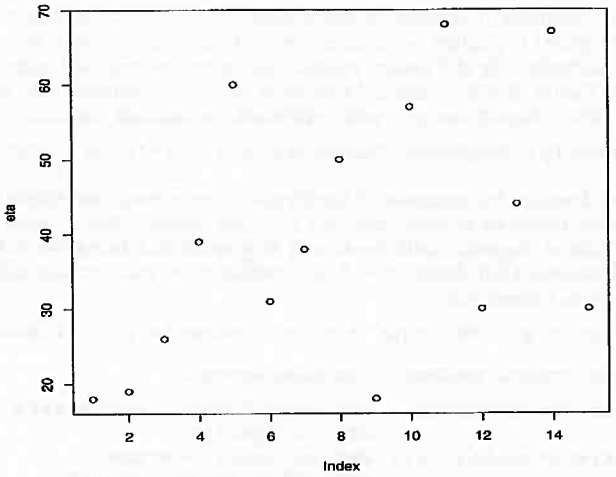

FIGURA R2.1. *Rappresentazione grafica elementare di una sequenza di dati numerici.*

```
## conserva i precedenti valori in vecchipar ed attiva i nuovi
R==> vecchipar=par(col="red",lwd=4);
R==> ............................  ## istruzioni grafiche
## ripristina i parametri grafici originari
R==> par(vecchipar);
```

A tale proposito, qui ed ovunque nel seguito, il lettore dovrà tenere conto del fatto che la stampa in bianco e nero non riproduce i colori delle istruzioni (ma che vengono mantenute per consentire al lettore una migliore visualizzazione sullo schermo del suo PC).

Ora, in contrapposizione alla semplicità della Figura R 2.1 e per esemplificare talune opportunità grafiche di R , disegniamo nella Figura R 2.2 la funzione $y = e^x$ nell'intervallo $[0, 1]$ e poi sovrapponiamo polinomi di grado crescente (sovrapposte a quella originaria mediante la funzione curve(), che disegna una funzione di una sola variabile *sempre* indicata con x) per evidenziare come l'approssimazione migliori sensibilmente al crescere del grado del polinomio. Le opzioni scelte per costruire tale figura dovrebbero essere attentamente esaminate e sono agevoli da ricostruire, passo dopo passo, anche grazie ai commenti che precedono le istruzioni seguenti.

```
### Definizione ascisse e ordinate della funzione y=exp(x)
R==> x=seq(0,1,by=0.01); y=exp(x);

### Istruzione primaria per il grafico con ampiezza (lwd=10)
### e tonalità di grigio in modo da distinguerlo dalle altre linee
R==> plot(x,y,type="l",lwd=10,col=gray(0.8),cex=0.9,
       main="Approssimazioni mediante polinomi
       della funzione exp(x)");

### Aggiunta dei termini di 1°,2°,3°,4° grado, che R calcola
### per lo stesso intervallo di valori di x definito
### nell'istruzione primaria, e disegnati con differenti tipologia
```


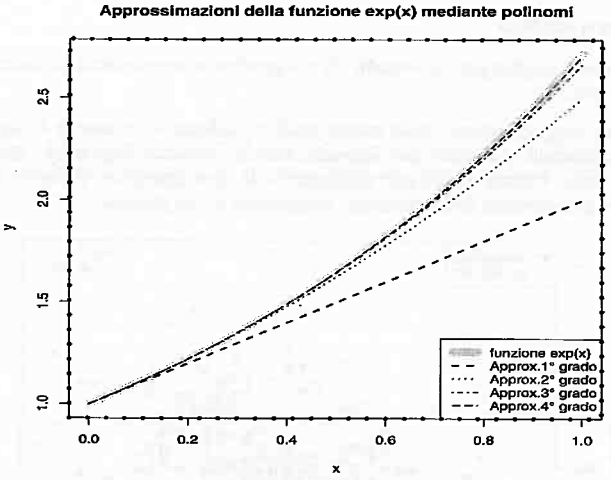

FIGURA R2.2. *Funzione esponenziale ed approssimazioni polinomiali.*

```
### delle linee (lty=2,3,4,5) ed ampiezze decrescenti
R==> curve(1+x,add=TRUE,lty=2,lwd=2);
R==> curve(1+x+(x^2)/2,add=TRUE,lty=3,lwd=1.75);
R==> curve(1+x+(x^2)/2+(x^3)/6,add=TRUE,lty=4,lwd=1.5);
R==> curve(1+x+(x^2)/2+(x^3)/6+(x^4)/24,add=TRUE,lty=5,lwd=1.25);

### La legenda è posizionata in basso a destra con tutte
### le opzioni delle funzioni precedenti
R==> legend("bottomright",legend=c("funzione exp(x)",
       "Approx.1° grado","Approx.2° grado",
       "Approx.3° grado","Approx.4° grado"),
       lwd=c(10,2,1.75,1.5,1.25),lty=1:5,
       col=c(gray(0.8),rep("black",4)));

### Infine, viene aggiunto al grafico un riquadro,
### che è anche colorato e tratteggiato
R==> box(lwd=5,lty=3,col="purple");
```

Il grafico si presenta in una propria finestra (in aggiunta a quella di R) ed è anche possibile interagire tramite il mouse; infatti, con i comandi in alto o con il tasto destro del mouse è possibile salvare il grafico in formati convenienti per altri software (per esempio, con l'estensione *eps*, preferibile per chi lavora in LaTeX). Chi lavora con *Word*, può posizionarsi sulla finestra grafica con il mouse, fare clic con il pulsante destro e scegliere l'opzione "Copia come metafile"; quindi, per inserire tale grafico nel documento di *Word* nel posto prescelto basterà applicare Ctrl+V e, poi, adattare le dimensioni del grafico alle esigenze e agli spazi del testo in *Word* mediante successive applicazioni del mouse. Infine, si può anche generare direttamente un file pdf contenente il grafico.

Interazione con il grafico

Alcune funzioni migliorano la lettura di un grafico e consentono di interagire con i valori rappresentati.

- Quando si rappresentano sullo stesso grafico differenti fenomeni è opportuno associare ai simboli utilizzati una legenda, con la funzione `legend()`, come esemplifichiamo nella Figura R 2.3 per distinguere le due tipologie di foglie nel data set `camellia` a proposito delle variabili `lunghezza` e `larghezza`:

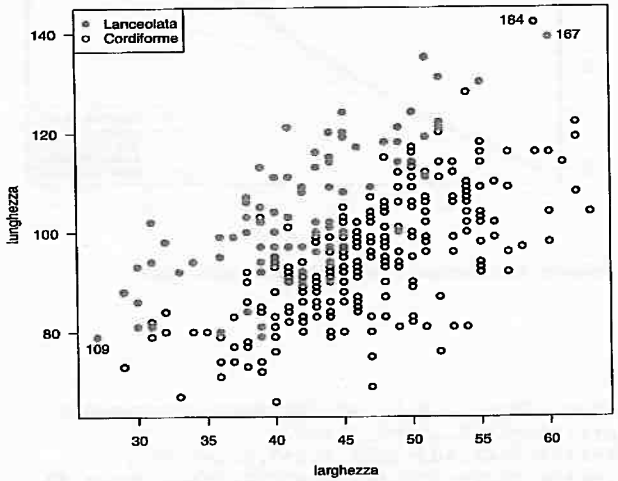

FIGURA R2.3. *Variabili lunghezza rispetto a larghezza, nel data set camellia.*

```
### Leggi CAMELLIA
R==> dati=read.table("camellia.txt", header=TRUE);
R==> lunghezza=dati[,2]; larghezza=dati[,3]; codice=dati[,6];
### codifica differenti colori in funzione della specie
R==> colspecie=ifelse(codice==1,gray(0.6),"black");
### codifica differenti simboli in funzione della specie
R==> simbspecie=ifelse(codice==1,19,21);
### simboli più marcati (lwd=2)
### numeri sugli assi in orizzontale (las=1)
R==> plot(larghezza,lunghezza,pch=simbspecie,
           col=colspecie,lwd=2,las=1);
### aggiungi legenda in alto a sinistra della figura
R==> legend("topleft",legend=c("Lanceolata","Cordiforme"),
       col=c(gray(0.6),"black"),pch=c(19,21));
```

La collocazione della legenda può variare qualificando in modo differente la prima stringa di caratteri (le opzioni sono su `?legend`), oppure fornendo le coordinate cartesiane, oppure con la funzione `locator` (descritta più avanti). Tra le istruzioni, si è utilizzata la funzione `ifelse(condizione,a,b)` che assume il valore `a` se `condizione` è vera, il valore `b` se `condizione` è falsa: in tal modo, si distinguono le due specie in termini di colore e simbolo.

- La funzione `identify()` consente di individuare (e segnare), sul grafico già disegnato mediante `plot()`, l'indice (o un nome, se è attivata tale opzione) di $n \geq 1$ punti mediante successivi clic del mouse sinistro. Le istruzioni seguenti segnano $n = 3$ punti sulla Figura R 2.3, utilizzando anche il colore; si evidenzia che essi sono il 109°, 167°, 184°, rispettivamente, così indicizzati nel data set `camellia`:

```
R==> identify(larghezza,lunghezza,n=3,col="blue",lwd=0.8);
```

- La funzione `locator(n)` consente di localizzare le coordinate cartesiane di n punti su un grafico esistente facendo clic con il mouse sinistro per n volte: alla fine, tali coordinate si ritrovano nella finestra di R (oppure in alternativa in una `lista` scelta dall'utente). Così, fissato $n = 3$, le coordinate dei punti prima individuati si localizzano con l'istruzione:

```
R==> locator(3); ### dopo 3 clic, coordinate in finestra di R
```

oppure, in alternativa, mediante le istruzioni seguenti:

```
R==> quali=locator(3);    ### dopo 3 clic, coordinate
                          ### in "quali"
R==> ascisse=quali[[1]]; ### recupera ascisse
                          ### dalla lista "quali"
R==> round(ascisse,digits=0); ### arrotonda all'intero
                               ### più vicino
[1] 27 60 59
R==> ordinate=quali[[2]]; ### recupera ordinate
                          ### dalla lista "quali"
R==> round(ordinate,digits=0); ### arrotonda
                               ### all'intero più vicino
[1]  79 139 142
```

Poche considerazioni per chiarire alcuni comandi finali: l'oggetto `quali` è della classe *list* (una categoria accennata alla fine del paragrafo 2.3 e che sarà brevemente ripresa nel paragrafo 2.12) per cui, per recuperarne gli elementi, occorrono le doppie parentesi quadre; inoltre, poiché la localizzazione del mouse varia impercettibilmente nei dintorni del punto prescelto ad ogni clic, è buona norma arrotondare il risultato (come effettuato nelle istruzioni precedenti, richiedendo 0 decimali). Infine, il posizionamento dei numeri sulla figura si colloca attorno al punto prescelto con il mouse per cui spostandosi impercettibilmente si ottimizza la visualizzazione.

Le funzioni `legend()`, `identify()` e `locator()` si utilizzano indipendentemente tra loro, purché su un grafico già creato mediante la funzione `plot()`; ma si possono anche utilizzare congiuntamente, per esempio, per collocare con il mouse la legenda in un punto prescelto, come con l'istruzione seguente:

```
R==> legend(locator(1),legend=c("Lanceolata","Cordiforme"),
     col=c(gray(0.6),"black"),pch=c(19,21));
```

che consente di collocare la medesima legenda di cui alla Figura R 2.3 nel punto dove l'utente effettua clic con il mouse sinistro.

R2.10 Rappresentazioni di variabili qualitative

Se si dispone di variabili qualitative, un grafico molto diffuso è quello a torta (*pie chart*) che presuppone la disponibilità dei dati in forma tale che a fronte di ciascuna modalità qualitativa vi sia il corrispondente ammontare (*frequenza*).

Per esempio, dal data set `studenti` rappresentiamo la distribuzione di frequenza degli studenti rispetto al diploma secondario. Le colonne 6-9 di tale data set contengono le variabili dicotomiche `sci`, `cla`, `tec`, `alt`, che assumono i valori 0 ed 1 a seconda che il diploma in esame sia posseduto oppure no dallo studente; inoltre, nella colonna 10, si dispone della variabile qualitativa (=*factor*) `diploma` che presenta una delle quattro modalità (=*levels*) `"sci"`, `"cla"`, `"tec"`, `"alt"` per ciascuno studente. Quindi, l'insieme delle istruzioni per ottenere la Figura R 2.4 sono le seguenti:

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> tipodip=c("altri","classico","scientifico","tecnico");
R==> with(data=dati, pie(table(diploma),tipodip,font=4,lwd=3,
       col=gray(c(0.2,0.4,0.6,0.8))));
```

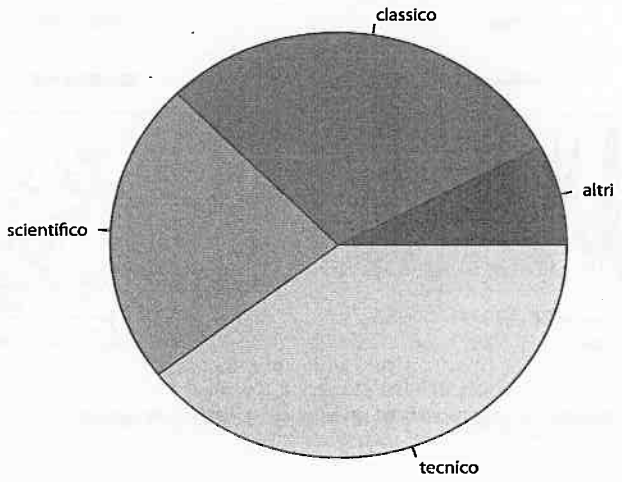

FIGURA R2.4. *Rappresentazione a torta della variabile `diploma` nel data set `studenti`.*

Come vedremo nel Capitolo 3, la funzione `table()` genera le frequenze assolute di una variabile qualitativa, associando la denominazione delle modalità qualitative presenti nel file (come è evidente dalla Figura R 2.4). Inoltre, abbiamo utilizzato la funzione `with()` che permette l'uso temporaneo delle variabili presenti nel dataset, senza la necessità di usare `attach()`. Poi, l'opzione `tipodip` permette si scrivere le modalità attorno ai settori della torta (si osservi che l'ordinamento delle modalità è quello alfabetico); inoltre, abbiamo utilizzato la funzione `gray()` per tenere conto di sfumature di grigio nella stampa in bianco e nero (ma si possono usare molte altre opzioni per i colori). Infine, nel dimensionare la stampa del grafico, occorre porre attenzione affinché l'area resti quadrata per non deformare la torta.

Tuttavia, dobbiamo sottolineare che gli statistici diffidano dall'usare tali rappresentazioni in modo categorico. A tal fine, giova citare testualmente le note (riportate in molti libri di Statistica) che accompagnano l'uso della funzione `pie()` nell'aiuto di R :

Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data. Cleveland (1985), page 264: "Data that can be shown by pie charts always can be shown by a dot chart. This means that judgements of position along a common scale can be made instead of the less accurate angle judgements." This statement is based on the empirical investigations of Cleveland and McGill as well as investigations by perceptual psychologists.

Il lettore può giudicare da sé quanto elevata sia la qualità scientifica delle informazioni quotidianamente diffuse dai mass-media le quali, invece, abbondano ed abusano regolarmente di tali rappresentazioni grafiche...

Invece, occorre preferire decisamente il *diagramma a barre* il quale, nel caso precedente e con le stesse informazioni, si attiva mediante le istruzioni seguenti:

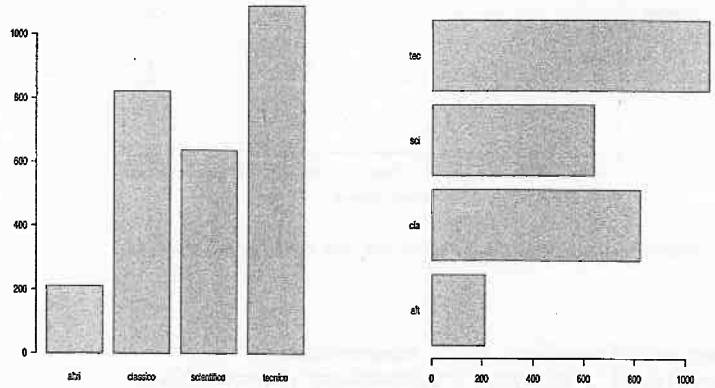

FIGURA R2.5. *Diagramma a barre della variabile `diploma` nel data set `studenti` (in orizzontale e verticale).*

```
R==>dati=read.table("studenti.txt", header=TRUE);
R==> tipodip=c("altri","classico","scientifico","tecnico");
### Figura di sinistra
R==> barplot(table(dati$diploma), las=1,names.arg=tipodip);
### Figura di destra
R==> barplot(table(dati$diploma), horiz=TRUE,las=1);
```

La Figura R 2.5 (a sinistra, con barre verticali, e a destra, con barre orizzontali) mostra che la rappresentazione grafica delle medesime informazioni mediante il grafico a barre è decisamente più utile per offrire interpretazioni immediate ed univoche. L'opzione `las=1` (comune a molte funzioni grafiche primarie) permette di scrivere il nome delle etichette sulle ordinate in orizzontale (cioè perpendicolarmente rispetto all'asse). Ulteriori opzioni per la funzione `barplot` sono possibili (per titoli, etichette sugli assi, colori, spazi tra le barre, e così via).

Infine, citiamo una variante orizzontale del diagramma a barre, cioè il *grafico a punti* di Cleveland che si ottiene tramite la funzione `dotchart`. Esso non è stato esaminato nel

testo, per cui ci limitiamo a presentare nella Figura R 2.6 una esemplificazione ottenuta con gli stessi dati di cui alla Figura R 2.5.

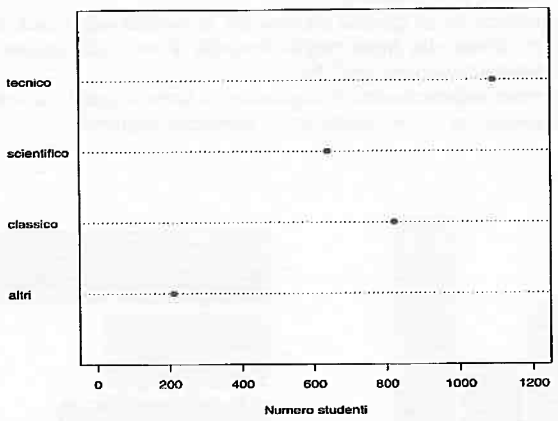

FIGURA R2.6. *Diagramma a punti della variabile diploma nel data set studenti.*

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> tipodip=c("altri","classico","scientifico","tecnico");
R==> dotchart(table(dati$diploma),labels=tipodip,pch=19,
              xlab="Numero studenti",xlim=c(0,1200));
```

R2.11 Rappresentazioni di variabili quantitative

Nel caso di variabili quantitative, la funzione `plot` esegue la rappresentazione dei grafici di una sequenza di modalità senza alcuna altra caratteristica. Le istruzioni successive (applicate ad un caso ove non è particolamente utile la rappresentazione prescelta) evidenziano come con le opzioni della funzione `plot` si possono ottenere grafici differenti in termini di simboli, tipologia e intensità delle linee, etichette sugli assi, titoli, etc. (escludiamo le opzioni sui colori non essendo ottimali per una stampa in bianco e nero).

Ciò sarà mostrato suggerendo al lettore di implementare il codice seguente e di studiare attentamente gli esiti che si ottengono, sugli stessi dati, rappresentati in modo differente nei 4 grafici successivi. La suddivisione della finestra grafica in 2 grafici per riga per 2 colonne si ottiene con la funzione `par(mfrow=c(2,2))`; alla fine, però, occorre sempre ripristinare la usuale rappresentazione di 1 grafico in 1 solo quadrante inserendo l'istruzione `par(mfrow=c(1,1))`.

Il lettore confrontando le istruzioni nei 4 differenti riquadri dovrebbe riuscire ad individuare agevolmente gli effetti indotti dalle differenti opzioni utilizzate: una completa elencazione si ha richiedendo aiuto con `?par()`; in effetti, la funzione `par()` contiene quasi tutti i parametri utili per adattare il grafico alle proprie esigenze. Si osservi, poi, che talvolta il titolo principale (`main=...`) è stato inserito nella funzione grafica primaria `plot()` e

talvolta separatamente, dopo `plot()`, utilizzando la funzione grafica secondaria `title()`.

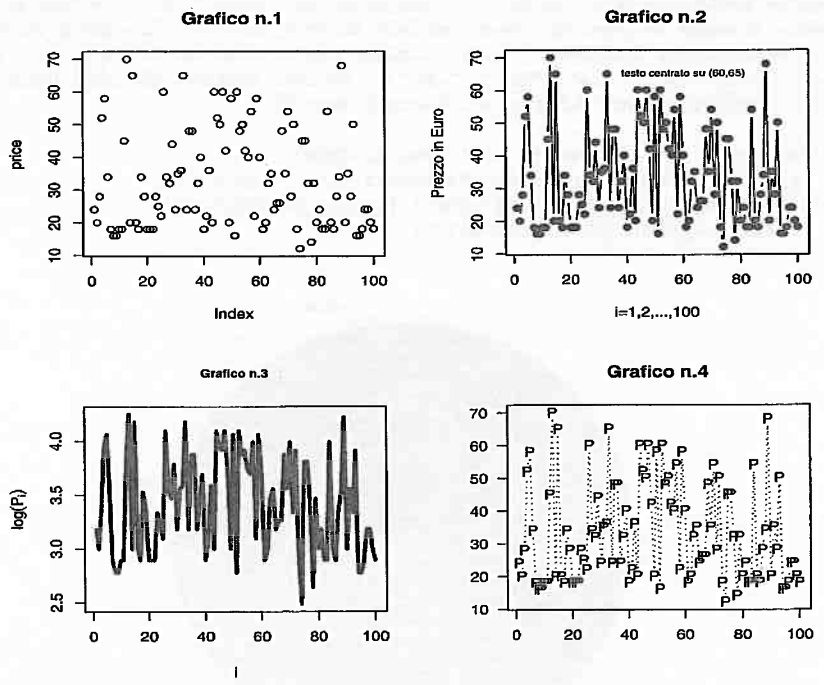

FIGURA R2.7. *Esempio di utilizzo della funzione `plot` con differenti opzioni.*

```
R==> dati=read.table("prezzolibri.txt", header=TRUE); ## legge
                                                      ## prezzi
R==> price=dati[1:100,]   # price contiene i primi 100 dati

R==> par(mfrow=c(2,2))  # suddividi area grafica in 4 regioni

### Grafico 1, con opzioni standard minimali
### e con aggiunta separata del titolo
R==> plot(price);
R==> title(main="Grafico n.1");

### Grafico n.2, grafico con testo aggiunto
R==> plot(price, xlab="i=1,2,...,100", ylab= "Prezzo in Euro",
          pch=19, type="b", main="Grafico n.2");
R==> text(60,65,cex=0.7,"testo centrato su (60,65)");

### Grafico n.3, log(prezzi), etichette
### con espressioni matematiche
```

```
### linee unite e spesse
R==> plot(log(price),xlab=expression(i),
         ylab=expression(log(P[i])),pch=1, type="l",lwd=3);
R==> title("Grafico n.3", cex.main=0.8),    # riduci all'80%
                                            # il titolo
R==> text(10,40,"log(prezzi)");

### Grafico n.4, differenti opzioni
### simbolo "P", punti uniti con linee punteggiate
### titolo ingrandito del 20%, etichette assi in orizzontale
R==> plot(price,xlab="", ylab="", pch="P", lty="dotted",type="b",
         main="Grafico n.4", cex.main=1.2, las=1);

R==> par(mfrow=c(1,1));  # ripristina l'area del grafico
```

Il risultato produce la Figura R 2.7. Una volta generata nella finestra grafica di R , essa può essere salvata cliccando in alto nella finestra grafica su "File" (e poi sul formato desiderato).

R2.12 Cenni alle liste

Le liste sono una collezione di oggetti di R di natura diversa e, per un loro uso efficiente, si richiedono altre nozioni, non funzionali a questo testo. Ci limitiamo ad indicare alcune utilizzazioni che, di volta in volta, saranno fatte.

In primo luogo, una lista si crea nel modo seguente:

```
R==> mialista=list(c(a,b,c), 1:20, c(TRUE,FALSE;FALSE;TRUE));
R==> mialista;
[[1]]
[1] -1  2  4

[[2]]
[1] "A" "B" "C" "D" "E"

[[3]]
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20

[[4]]
[1]  TRUE FALSE FALSE  TRUE
```

Come si vede, la lista contiene elementi di qualsiasi natura (numeri, stringhe, variabili logiche, etc.) ed ogni oggetto è caratterizzato da `mialista[[1]]`, `mialista[[2]]`, `mialista[[3]]`, etc. Invece, gli elementi dei singoli oggetti vengono richiamati con le usuali parentesi quadre (e con le regole di indicizzazione già viste in precedenza), come nell'esempio seguente:

```
R==> mialista[[1]][3];
[1] 4
R==> mialista[[2]][3:5];
[1] "C" "D" "E"
R==> mialista[[3]][-(4:20)];
[1] 1 2 3
R==> mialista[[4]];
[1]  TRUE FALSE FALSE  TRUE
```


Nell'esempio, non sono stati assegnati dei nomi agli oggetti della lista. Ciò avviene mediante:

```
names(mialista)=c("primo","secondo","terzo","quarto");
```

Allora, gli oggetti possono essere richiamati nel modo usuale. Per esempio:

```
R==> mialista$quarto;
[1]  TRUE FALSE FALSE  TRUE
R==> mialista$secondo[2:4];
[1] "B" "C" "D"
```

R2.13 Rappresentazioni di serie storiche

La funzione `plot` potrebbe essere utilizzata anche per il grafico di serie storiche, essendo queste delle rilevazioni di una variabile ordinata rispetto al tempo; tuttavia, la specificità di dati temporali suggerisce un trattamento più idoneo.

In primo luogo, occorre convertire il vettore di osservazioni (che chiameremo `serie`) in un oggetto della classe *time series* stabilendo la data iniziale e la frequenza di rilevazione rispetto all'anno (= 12, per dati mensili; = 4, per dati trimestrali, etc.). Quindi, si attiva l'istruzione `plot`.

Esemplifichiamo tali procedure sulla serie storica dell'indice generale della produzione industriale in Italia (misurato su base mensile), presente nel data set `produzione`. Le istruzioni che seguono mirano a fornire, in una sola rappresentazione grafica, diversi aspetti di natura esplorativa di una serie storica, sui quali non ci soffermiamo.

A tale scopo, molte delle istruzioni in R che presentiamo sono finalizzate alla personalizzazione dei grafici, e sono immediatamente comprensibili. Si noti che ciò si semplifica se si attiva un ciclo (con il comando `for`, che esamineremo in dettaglio nel paragrafo 6.5) il quale, per ciascun anno, disegna la linea di interesse; a tal fine, è opportuno definire prima il riquadro di riferimento (mediante la funzione `plot` e l'opzione `type=''n''` che non disegna alcunché), all'interno del quale poi le singole linee vengono disegnate con la funzione `lines()` (che è una funzione grafica secondaria). Infine, per ottenere il grafico presentato nel quarto quadrante, occorre introdurre ed eseguire la funzione `polare()`, qui implementata per offrire una importante caratterizzazione di serie stagionali mensili: un'attenta lettura di tale funzione consentirà al lettore di acquisire una certa competenza per costruire da sé funzioni utili alle sue specifiche elaborazioni.

```
R==> polare<-function(serie){
        ciclo=12;
        punti=480;xp=2*pi*(0:punti)/punti;
        pseni=sin(xp); pcoseni=cos(xp);
        n=length(serie);
        succ=2*pi*(0:(n-1))/ciclo;
        seni=sin(succ); coseni=cos(succ);
        minser=min(serie); maxser=max(serie);
        seriestar=(serie-minser)/(maxser-minser); ### maps to
                                                  ### [0,+1]
        ascisse=seriestar*coseni;
        ordinate=seriestar*seni;
        nomemesi=c("Gen","Feb","Mar","Apr","Mag","Giu",
            "Lug","Ago","Set","Ott","Nov","Dic");
        dovenomi=c(rep(4,2),rep(3,3),rep(2,4),1,rep(4,2));
        offset=0.5;
```

```
    plot(coseni,seni,xlim=c(-1.2,1.2),ylim=c(-1.2,1.2),
        xlab="",ylab="",pch=5,col="red",lwd=2,pty="s");
    text(coseni,seni,labels=nomemesi,pos=dovenomi,
        font=4,col="red");
    lines(pseni,pcoseni,lty=2);
    segments(0,0,seni,coseni,lwd=0.7,lty=6);
    lines(ascisse,ordinate,pch=19,type="b",lwd=1);
}
```

In sequenza, nella Figura R 2.8 sono rappresentate la successione temporale dell'indice mensile della produzione industriale, la stessa serie suddivisa per anni (allo scopo di far emergere il pattern stagionale) e poi per ciascun mese di anni differenti (per evidenziare eventuali andamenti tendenziali) e, infine, la rappresentazione polare (per far emergere ripetizioni di comportamento nei vari anni).

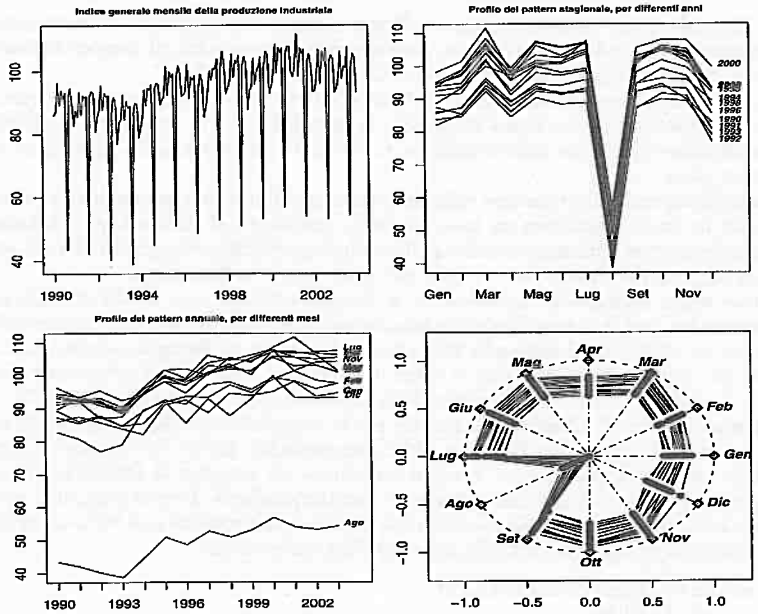

FIGURA R2.8. *Rappresentazioni grafiche per l'indice mensile della produzione industriale.*

```
R==> dati=read.table("produzione.txt", header=TRUE);
R==> produzione=dati[,1];    anni=length(produzione)/12;

R==> anninomi=as.character(seq(1990,2003,by=1));
R==> mesinomi=c("Gen","Feb","Mar","Apr","Mag","Giu",
          "Lug","Ago","Set","Ott","Nov","Dic");
R==> shifttitle=0.7;

R==> par(mfrow=c(2,2))    ### suddividi area grafica in 4 regioni
R==> par(mar=c(2,2,2,1)+0.1) ### fissa nuovi margini
```



```
### Grafico 1: serie rispetto al tempo
R==> serie=ts(produzione,start=c(1990,1),frequency=12);
R==> plot(serie, xlab="t",ylab="",cex.main=0.7,
        main="Indice generale mensile
        della produzione industriale");
### Grafico n.2: profilo stagionale, per anni differenti
R==> plot(1:13,seq(40,110,length.out=13),xlab="",
        ylab="Serie x_t per anno",xaxt="n",type="n");
R==> title(main="Profilo del pattern stagionale,
        per differenti anni",
        cex.main=0.7);
R==> axis(1,at=1:12,labels=mesinomi);
        for(i in 1:anni){
            inizio=12*(i-1)+1; fine=inizio+11;
            prodmanno=produzione[inizio:fine];
            lines(1:12,y=prodmanno,type="l");
            text(12+shifttitle,produzione[fine]+shifttitle,
                anninomi[i],cex=0.7,font=4);
    }
### Grafico n.3: profili tendenziali, per mesi differenti
R==> plot(1:15,seq(40,110,length.out=15),xlab="",
        ylab="Serie x_t per mese",xaxt="n",type="n");
R==> title(main="Profilo del pattern annuale,
        per differenti mesi",
        cex.main=0.7);
R==> axis(1,at=1:14,labels=anninomi);
        for(i in 1:12){
            inizio=i; fine=inizio+12*(anni-1);
            indici=seq(inizio,fine,by=12);
            prodpermese=produzione[indici];
            lines(1:anni,y=prodpermese,type="l");
            text(14+shifttitle,produzione[fine]+shifttitle,
                mesinomi[i],cex=0.7,font=4);
    }
### Grafico n.4: grafico polare
R==> polare(produzione);

R==> par(mar=c(5,4,4,2)+0.1);    ###  ripristina margini
R==> par(mfrow=c(1,1));          ### ripristina area grafica
```

La Figura R 2.8 conferma il carattere fortemente stagionale della serie produzione, ben evidente nel picco in discesa di Agosto, ed un ciclo-trend leggermente crescente negli anni indicati.

Il carattere dinamico delle serie storiche richiede l'analisi delle variazioni rispetto al tempo e, tra queste, per dati mensili, sono importanti le differenze successive: quelle *ordinarie*, ottenute da x_t mediante $x_t - x_{t-1}$, e quelle *stagionali*, ottenute da x_t mediante $x_t - x_{t-12}$. Le variazioni ordinarie di ciascun mese rispetto al precedente sono dette *congiunturali*, quelle stagionali di ciascun mese rispetto allo spesso mese dell'anno precedente sono dette *tendenziali*; dal punto di vista statistico, è utile considerare anche le differenze ordinarie di quelle stagionali. Alla fine, per una serie storica mensile è importante visualizzare tali differenze e la corrispondente distribuzione di frequenza (di cui parleremo nel Capitolo 3).

Tali calcoli sono immediati in R mediante la funzione diff() che si può utilizzare

nelle due opzioni qui evidenziate:

```
R==> d1=diff(serie);          ### calcola d1(t)=x(t)-x(t-1)
R==> d12=diff(serie,lag=12); ### calcola d12(t)=x(t)-x(t-12)
R==> d1d12=diff(d12);         ### calcola d1d12(t)=d12(t)-d12(t-1)
```

Si noti che se la variabile serie consiste di n osservazioni, d1 ha $(n-1)$ osservazioni, d12 ne ha $(n-12)$ e d1d12 ne ha $(n-13)$. Le istruzioni seguenti generano tali rappresenta-

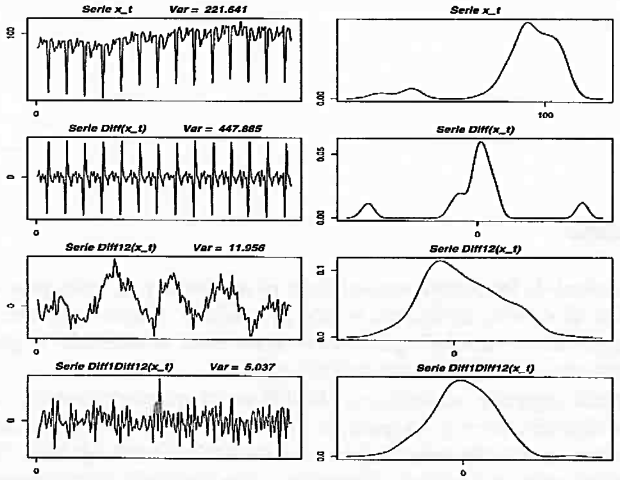

FIGURA R2.9. *Indice mensile della produzione industriale in Italia: serie originale, differenze successive e rispettivi istogrammi perequati.*

zioni e la successiva Figura R 2.9 visualizza i risultati per la serie mensile produzione nel data set produzione. Per la programmazione in R , è interessante osservare come – dovendo replicare le medesime istruzioni su oggetti differenti (le serie differenziate hanno differenti numerosità) – si crei la lista tutteserie e nel *loop* successivo ciascuna serie viene richiamata come elemento della lista mediante l'istruzione tutteserie[[j]], per i valori di $j = 1, 2, 3, 4$.

```
### Leggi PRODUZIONE
R==> dati=read.table("produzione.txt", header=TRUE);
R==> serie=dati[,1];              d1=diff(serie);
R==> d12=diff(serie,lag=12);      d1d12=diff(d12);
R==> tutteserie=list(serie,d1,d12,d1d12);
R==> titoli=c("Serie x_t","Serie Diff(x_t)",
       "Serie Diff12(x_t)","Serie Diff1Diff12(x_t)");
R==> par(mfrow=c(4,2));          ### suddividi l'area in 8 regioni
R==> par(mar=c(2,2,2,1)+0.1); ### fissa nuovi margini
       for(j in 1:4){
           xt=tutteserie[[j]];var=mean(xt^2)-(mean(xt))^2;
           plot(1:length(xt),xt,xlab="",
               ylab="",type="l",lab=c(1,1,0));
           axis(1,at=NULL,labels=FALSE,tick=FALSE);
           title(main=paste(titoli[j],"          Var = ",
                 round(var,3)),font.main=4);
           plot(density(xt),main=titoli[j],xlab="",
           ylab="",lab=c(1,1,0),font.main=4);
   }
R==> par(mar=c(5,4,4,2)+0.1);    ### ripristina margini
R==> par(mfrow=c(1,1));          ### ripristina l'area grafica
```


Dai grafici emerge come la dinamica delle variazioni sia molto diversa da quella della serie originaria e, soprattutto, come ciò modifichi la forma della distribuzione dei dati. Inoltre, va tenuto presente che cambia in modo notevole la variabilità delle serie considerate, come evidenziato dall'indice Var (il cui significato sarà esplicitato nel Capitolo 5).

Capitolo terzo

Distribuzioni di frequenza

3.1 Introduzione

Le distribuzioni di frequenza consentono di analizzare la gran parte dei fenomeni reali a fini di sintesi, confronto e interpretazione. Come vedremo, esse possono essere applicate a variabili qualitative (nominali o ordinali) o quantitative (discrete o continue) con opportune specificazioni.

Nel paragrafo seguente esaminiamo le differenti rappresentazioni tabellari e grafiche di una distribuzione di frequenza, in funzione della differente natura delle variabili rilevate e mediante esemplificazioni corrispondenti. Quindi, formalizzeremo tali risultati, con particolare riferimento alle variabili quantitative.

3.2 Rappresentazioni tabellari e grafiche

Anche se nella letteratura si distingue tra *distribuzione di quantità* (quando l'ammontare trasferibile di una variabile si ripartisce tra le unità statistiche della popolazione) e *distribuzione di frequenza*, noi ci limiteremo a studiare solo le seconde a causa della loro diffusione e rilevanza per la metodologia statistica.

Supponiamo, quindi, di aver osservato o rilevato sulle n unità statistiche di una popolazione ben definita una variabile X le cui modalità risultano essere: $x_1, x_2, ..., x_n$. Tale insieme di dati, pur essendo una informazione esaustiva sul fenomeno, può essere organizzato in forma tabellare per esplicitare – con maggiore immediatezza – taluni aspetti della variabile: minimo, massimo, modalità più frequente, valutazioni sintetiche, sottogruppi, etc.

Tale necessità diventa pressante quando la popolazione consiste di migliaia di unità statistiche per le quali la visualizzazione delle singole modalità rilevate non assume alcuna utilità decisionale per i compiti propri della Statistica: sintesi, confronto, interpretazione, evidenza empirica per le scelte, verifica di ipotesi, ricerca di leggi di comportamento, etc.

- La ***distribuzione di frequenza*** è una organizzazione dei dati in forma tabellare che ad ogni modalità di una certa variabile (qualitativa o quantitativa) fa corrispondere la rispettiva frequenza.

All'interno di una data popolazione, la distribuzione delle frequenze di una variabile evidenzia come il fenomeno si manifesta in rapporto alle differenti modalità di quella variabile. Se si esamina il numero di volte con cui si manifestano le differenti modalità si parla di ***frequenza assoluta***; se, invece, la frequenza assoluta è rapportata al totale delle unità statistiche della popolazione si parla di ***frequenza relativa***; infine, se la frequenza relativa è moltiplicata per 100 si parla di ***frequenza percentuale***.

La distribuzione di frequenza di una variabile quantitativa (che si presenta con modalità distinte) o di una variabile qualitativa (che si presenta con attributi distinti) è semplicemente un modo di organizzare e, quindi, di presentare le unità statistiche che manifestano la stessa modalità o lo stesso attributo sull'intera popolazione, enumerando le osservazioni per modalità o attributi coincidenti. Poiché lo studio coinvolge le frequenze (che sono numeri interi), è sempre possibile pervenire ad una distribuzione di frequenza per qualsiasi fenomeno.

In sostanza, la distribuzione di frequenza esplicita quante volte (in assoluto se trattasi di frequenza assoluta, oppure in rapporto al totale se trattasi di frequenza relativa) una determinata modalità si presenta nel collettivo in esame.

La rappresentazione grafica che pone in ascissa le modalità della variabile e in ordinata le frequenze assolute (o relative) si chiama ***diagramma a barre***, ampiamente discusso nel paragrafo R 2.10.

Gli esempi seguenti mostrano situazioni reali di distribuzioni di frequenza per differenti tipologie di variabili statistiche.

Esempio 3.1

∘ Dal data set STUDENTI, deriviamo la distribuzione della variabile qualitativa X=«Diploma di scuola secondaria», i cui possibili attributi sono riassunti nella tabella seguente (e rappresentati graficamente nella figura R 2.5):

Diploma	Frequenze		
	assolute	relative	percentuali
Classico	821	0.2976	29.76
Scientifico	637	0.2309	23.09
Tecnico	1090	0.3950	39.50
Altri	211	0.0765	7.65
Totali	2759	1.0000	100

La tabella mostra che 821 è la numerosità (frequenza assoluta) degli studenti/studentesse con Diploma secondario di tipo Classico. Tale quota esprime una frazione pari a 0.2976 del totale, cioè in termini percentuali il 29.76% ovvero poco meno del 30% delle unità statistiche considerate. Si comprende come i rapporti di composizione di questo esempio non siano altro che le frequenze relative delle differenti modalità.

□

Esempio 3.2

∘ I dati della tabella seguente mostrano la distribuzione di frequenza della *variabile qualitativa ordinale* «Giorno della settimana in cui preferibilmente vado al cinema» rilevato su una popolazione di 250 frequentatori abituali di sale cinematografiche.

Giorno della settimana	Frequenze		
	assolute	relative	percentuali
Lunedì	13	0.052	5.2
Martedì	15	0.060	6.0
Mercoledì	30	0.120	12.0
Giovedì	31	0.124	12.4
Venerdì	22	0.088	8.8
Sabato	92	0.368	36.8
Domenica	47	0.188	18.8
Totali	250	1.000	100.0

La rappresentazione grafica per questa distribuzione di frequenza può avvenire secondo le regole espresse per il diagramma a barre, ma il vincolo di ordine che sussiste tra le modalità suggerisce di applicare un diagramma polare (discusso nel paragrafo R 2.13).

□

Esempio 3.3

∘ I dati della tabella seguente concernono la distribuzione di frequenza della *variabile quantitativa discreta* (=«Numero dei componenti della famiglia») rilevata sui residenti in Lombardia e in Calabria, rispettivamente, durante il Censimento del 1981 della popolazione residente.

Numero di componenti	Lombardia		Calabria	
	Frequenze			
	assolute	relative	assolute	relative
1	591.927	0.1921	105.823	0.1700
2	743.032	0.2411	132.552	0.2130
3	747.740	0.2426	111.192	0.1787
4	665.163	0.2158	129.278	0.2077
5	233.871	0.0759	79.952	0.1285
6	100.054	0.0325	63.516	0.1021
Totali	3.081.787	1.0000	622.313	1.0000

La tabella (nella quale la modalità «6 componenti» racchiude in effetti le modalità «6 e più componenti») include due distinte distribuzioni di frequenze, per Lombardia e Calabria, rispettivamente, e consente una comparazione immediata tra la composizione dei nuclei familiari in queste due Regioni tramite il confronto delle corrispondenti frequenze relative.

Infatti, non è corretto confrontare le 132.552 famiglie della Calabria formate da 2 componenti con le corrispondenti 743.032 famiglie della Lombardia, perché il complesso delle famiglie calabresi è stato nel 1981 di 622.313 a fronte delle 3.081.787 famiglie lombarde.

È, quindi, coerente effettuare dei confronti relativi (o percentuali, se i rapporti vengono moltiplicati per 100) affermando che in Calabria le famiglie con 2 componenti costituiscono il 21.30% (= 100*132.552/622.313) mentre in Lombardia costituiscono il 24.11% (= 100*743.032/3.081.787): quindi, «relativamente», le famiglie con 2 componenti sono leggermente più numerose in Lombardia che in Calabria.

Inoltre, in Calabria le famiglie più diffuse erano quelle con 2 componenti (probabilmente, per l'emigrazione che provoca nei residenti una accentuazione della popolazione più anziana) mentre in Lombardia erano quelle con 3 componenti. Invece, con riferimento ai nuclei familiari più numerosi, la tabella evidenzia come – alla data indicata – il 23% delle famiglie in Calabria aveva 5 o più componenti, mentre in Lombardia a tali modalità corrispondevano poco più del 10% delle famiglie.

Per questo tipo di confronti, la rappresentazione grafica più idonea resta il diagramma a barre (Figura 3.1) che fornisce una buona comparazione di come si distribuisce la variabile «Numero di componenti» nelle due popolazioni.

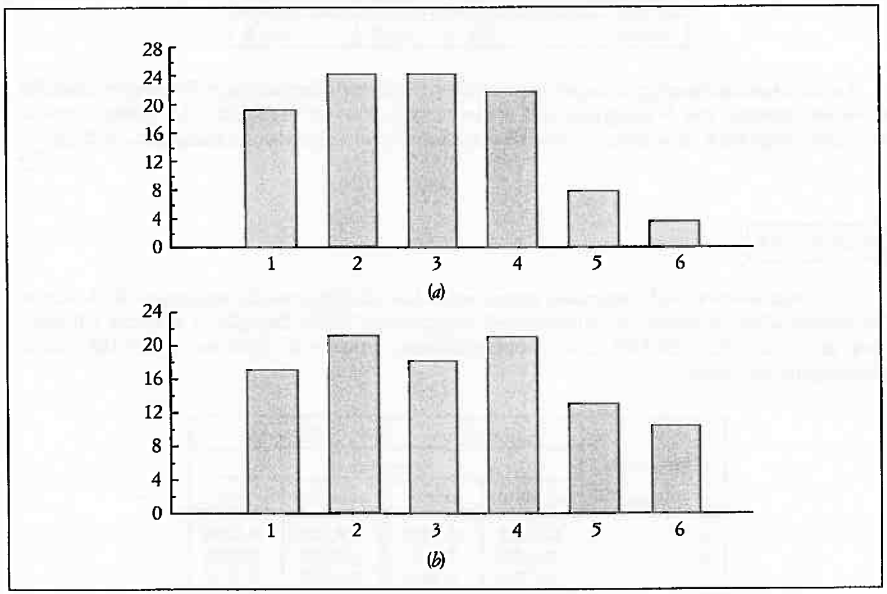

FIGURA 3.1. Diagramma a barre della variabile «Numero di componenti»: *a*) Lombardia; *b*) Calabria.

□

Esempio 3.4

○ I dati della tabella seguente (dal data set COMPACT) riguardano la distribuzione di frequenza della *variabile quantitativa continua* X = «Durata in secondi» per $n = 1192$ brani musicali. Il valore minimo è $x_{min} = 30$, il valore massimo è $x_{max} = 1022$ ma la gran parte dei brani presenta una durata non eccessiva.

Infatti, ben 826 brani su 1192 (cioè quasi il 70%) hanno durate comprese tra 120 e 300 secondi (cioè tra 2 e 5 minuti) e addirittura 1138 brani (pari al 95.5%) hanno durate comprese tra 120 e 600 secondi (cioè tra 2 e 10 minuti).

Per la natura della variabile appare ragionevole suddividere l'ampiezza dei valori osservati in classi che siano multiple di 60, in modo da interpretare immediatamente le informazioni derivate dalla tabella in minuti di durata del brano.

Classi di modalità	Frequenze assolute	Frequenze relative
$0 \vdash - \dashv 60$	5	0.00419
$60 - \dashv 120$	27	0.02265
$120 - \dashv 180$	255	0.21393
$180 - \dashv 300$	571	0.47903
$300 - \dashv 600$	312	0.26174
$600 - \dashv 1022$	22	0.01846
Totali	1192	1.00000

Se si volesse applicare il sistema dei diagrammi a barre per la rappresentazione di questi dati si commetterebbe sicuramente un errore perché le classi non sono tutte della stessa ampiezza. Per esempio, i 27 brani della classe (60,120] non sono confrontabili con quelli della classe (600,1022] perché l'ampiezza di quest'ultima è molto più ampia della prima.

Una soluzione corretta è l'*istogramma* (che discuteremo con ampiezza nel paragrafo 3.4) nel quale le barre contigue si disegnano in modo che sia l'area – e non l'altezza – la quantità corrispondente alle frequenze assolute. La Figura 3.2 rispetta questo criterio e visualizza, quindi, correttamente la distribuzione della variabile continua che si sta studiando. Essa è stata ottenuta suddividendo le durate (comprese tra 30 e 1022 secondi) in 11 classi di eguale ampiezza. In tal modo, esaminando le durate dei 1192 brani e contando quanti di essi ricadono in ciascuna delle classi anzidette, si ottiene la seguente *distribuzione di frequenza in classi equi-ampie*:

Classi di modalità (equi-ampie)	Frequenze assolute	Frequenze relative
$30 \leq X \leq 120.182$	32	0.02685
$120.182 < X \leq 210.364$	390	0.32718
$210.364 < X \leq 300.545$	436	0.36577
$300.545 < X \leq 390.727$	236	0.19799
$390.727 < X \leq 480.909$	48	0.04027
$480.909 < X \leq 571.091$	25	0.02097
$571.091 < X \leq 661.273$	11	0.00923
$661.273 < X \leq 751.455$	6	0.00503
$751.455 < X \leq 841.636$	5	0.00419
$841.636 < X \leq 931.818$	2	0.00168
$931.818 < X \leq 1022$	1	0.00084
Totali	1192	1.00000

□

Gli ultimi esempi hanno mostrato come nel caso di variabili discrete sia possibile elencare tutte le modalità che può assumere la variabile, mentre per variabili

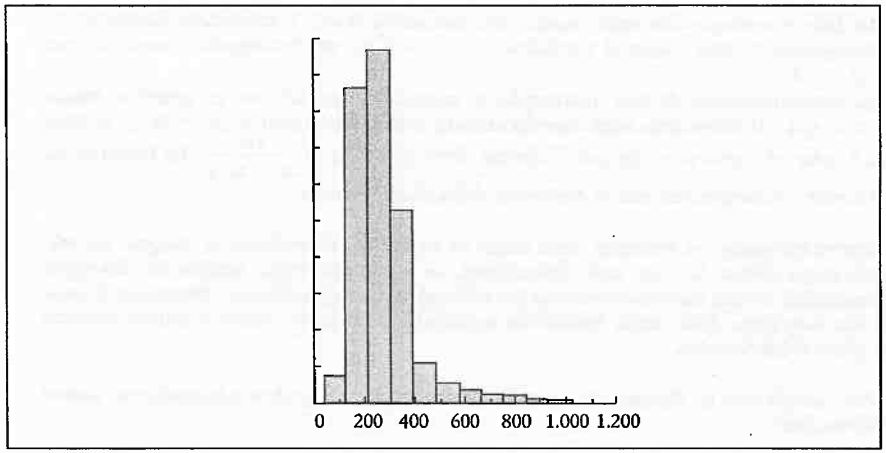

FIGURA 3.2. Istogramma della variabile «Durata dei brani» in classi equifrequenti.

continue occorrerà limitarsi a contare quante unità manifestano la variabile con modalità compresa in un certo intervallo di valori reali (il che induce una certa arbitrarietà, come vedremo nella definizione di tali classi). Pertanto, per variabili continue, occorre individuare sottoinsiemi dell'intervallo di definizione e contare la frequenza di tali sotto-intervalli piuttosto che la frequenza delle singole modalità.

In effetti, per fenomeni di tipo continuo (si pensi al peso di una foglia, al tempo necessario per raggiungere il luogo di lavoro ogni mattina, all'altezza di una persona adulta, alla durata di un film) la rilevazione darà luogo a numeri quasi sempre differenti tra loro. È solo per semplicità che diciamo che due foglie pesano 3 gr, che impieghiamo 40 minuti per il tragitto casa-lavoro, che due persone sono entrambe alte cm 172. In realtà, nelle due circostanze a confronto, le risposte coincidono solo perché ci riferiamo ad un valore approssimato determinato, rispettivamente, sulla bilancia, sull'orologio o sul metro.

Tale distinzione è molto importante per cui, nella successiva formalizzazione, è necessario distinguere accuratamente le distribuzioni di frequenza delle *variabili discrete* da quelle delle *variabili continue.*

3.3 Distribuzioni di frequenza per variabili discrete

Una popolazione è composta di n unità statistiche sulle quali si rileva una variabile quantitativa X che assume k modalità distinte $\{x_1, x_2, \ldots, x_i, \ldots, x_k\}$. Se la modalità x_1 si presenta n_1 volte, la modalità x_2 si presenta n_2 volte, ..., la modalità x_i si presenta n_i volte, ..., la modalità x_k si presenta n_k volte, allora è conveniente rappresentare queste informazioni in forma tabellare mediante la

seguente *distribuzione di frequenza* della variabile X:

Modalità della variabile X	Frequenze	
	assolute	relative
x_1	n_1	$f_1 = \frac{n_1}{n}$
x_2	n_2	$f_2 = \frac{n_2}{n}$
...	...	...
x_i	n_i	$f_i = \frac{n_i}{n}$
...	...	...
x_k	n_k	$f_k = \frac{n_k}{n}$
Totali	n	1

Le quantità n_i, $i = 1, 2, \ldots, k$ si definiscono ***frequenze assolute*** e sono sempre numeri interi non-negativi caratterizzati dalle proprietà:

$$n_1 + n_2 + \ldots + n_k = \sum_{i=1}^{k} n_i = n; \quad 0 \leq n_i \leq n, \quad i = 1, 2, \ldots, k.$$

Poiché le frequenze assolute dipendono dal totale n, per confrontare popolazioni differenti formate da un numero diverso di unità statistiche, è opportuno definire le ***frequenze relative***:

$$f_i = \frac{n_i}{n}, \; i = 1, 2, \ldots, k,$$

le cui proprietà si deducono da quelle delle frequenze assolute. Infatti, dividendo per n le precedenti relazioni sulle frequenze assolute, si ha:

$$f_1 + f_2 + \ldots + f_k = \sum_{i=1}^{k} f_i = 1; \quad 0 \leq f_i \leq 1, \; i = 1, 2, \ldots, k.$$

Gli esempi hanno mostrato come la rappresentazione grafica «a barre» verticali sia particolarmente idonea per la distribuzione di frequenza di tale tipologia di variabili.

3.4 Distribuzioni di frequenza per variabili continue

Nel caso di una variabile continua X non è possibile far corrispondere ai valori che essa assume le rispettive frequenze (assolute e relative), perché tra due modalità qualsiasi ve ne possono essere infinite altre, con la conseguenza che, quasi certamente, si avrebbe una sequenza di osservazioni distinte (cioè con frequenza assoluta pari ad 1) e poco distanti tra loro.

Appare, quindi, conveniente, per fornire una idea sintetica della distribuzione di una variabile continua, suddividere l'intervallo dei valori che X può assumere

(tra il minimo e il massimo) in classi di modalità (cioè sub-intervalli dell'intervallo di definizione) riferendo la distribuzione delle frequenze agli elementi che appartengono a ciascuna classe così considerata.

Per individuare criteri di suddivisione in classi ci si riferisce alle caratteristiche del fenomeno (come è stato fatto nell'esempio 3.4) oppure ad alcune convenzioni (per esempio, ai multipli di 5, 10, etc.); nel seguito, discuteremo alcune di tali possibili opzioni.

In ogni caso, quando per variabili continue si parla di frequenza (assoluta e relativa) con riferimento ai valori di una certa classe di modalità, e non più alla singola modalità, va inteso che essa indica quante unità statistiche presentano valori della variabile inclusi nell'intervallo della classe.

In pratica, se una generica classe viene definita come l'insieme dei valori inclusi nell'intervallo reale:

$$(x_{i-1},\, x_i] \;\; i = 1, 2, \ldots, k$$

in essa vanno conteggiate tutte le modalità di X comprese nell'intervallo reale: $x_{i-1} < X \leq x_i$, e la frequenza assoluta $n_i\,, i = 1, 2, \ldots, k$ indica quante unità statistiche presentano valori di X che ricadono in tale intervallo.

Pertanto, la riduzione dei dati in classi di modalità riduce l'informazione originaria a sotto-insiemi più gestibili ma, ovviamente, da una tale distribuzione di frequenza non sarà più possibile risalire alla rilevazione originaria.

Qualche accorgimento si rende necessario per individuare correttamente le classi estreme, anche sul piano della notazione. A tal fine, è conveniente indicare con $x_0 = x_{min} = min(X)$ il valore minimo di X e con $x_k = x_{max} = max(X)$, il valore massimo di X e poi suddividere l'intervallo $x_0 \leq X \leq x_k$ delle possibili modalità di X in k classi. La tabella seguente mostra lo schema standard mediante il quale si presenta una distribuzione di frequenza per variabili continue.

Per indicare tutti i valori di una variabile X compresi nell'intervallo da a (escluso) sino a b (incluso), cioè i valori tali che $a < X \leq b$ si utilizzano le scritture $(a, b]$ oppure $a \dashv b$; analogamente, per altri tipi di intervalli.

Gli intervalli che definiscono le classi sono in numero di k e sono tutti aperti a sinistra e chiusi a destra, eccetto il primo (per tener conto anche del minimo valore x_0 che viene incluso nella prima classe).

Classi di modalità della variabile X	Frequenze	
	assolute	relative
$x_0 \leq X \leq x_1$	n_1	$f_1 = \frac{n_1}{n}$
$x_1 < X \leq x_2$	n_2	$f_2 = \frac{n_2}{n}$
...	...	...
$x_{i-1} < X \leq x_i$	n_i	$f_i = \frac{n_i}{n}$
...	...	...
$x_{k-1} < X \leq x_k$	n_k	$f_k = \frac{n_k}{n}$
Totali	n	1

Poiché le classi possono avere ampiezza differente, non è corretto utilizzare la rappresentazione del diagramma a «barre» perché tali barre si collocherebbero su classi di ampiezza differente. Per questo, allo scopo di avere un impatto visivo coerente (che è

fornito dalla visualizzazione della consistenza numerica di n_i), è necessario disegnare, in corrispondenza di ogni classe di modalità $x_{i-1} < X \leq x_i$, un rettangolo di area n_i, per $i = 1, 2, \ldots, k$.

La determinazione di tale rettangolo è immediata poiché per la generica classe $x_{i-1} < X \leq x_i$, il rettangolo avrà base facilmente calcolabile e pari a: $(x_i - x_{i-1})$ mentre l'area è nota ed è pari a n_i. Quindi, l'altezza deve essere: $h_i = \dfrac{n_i}{x_i - x_{i-1}}$. La quantità h_i viene a volte chiamata *densità di frequenza* della classe i-esima.

Operativamente, su ciascuna delle classi di modalità individuate, si disegna un rettangolo la cui altezza h_i viene così determinata. La rappresentazione grafica che alle classi delle modalità di una variabile continua (in ascissa) fa corrispondere un rettangolo di area pari alla frequenza delle unità statistiche appartenenti a quella classe è universalmente nota come ***istogramma***.

Per completare la discussione su tale rappresentazione grafica aggiungiamo alcune considerazioni:

- La rappresentazione tramite istogramma è coerente con la rappresentazione fornita per le variabili discrete poiché, per queste ultime, si può pensare che la classe sia ridotta ad un solo punto e, quindi, l'area si riduca alla sola altezza pari ad n_i.
- La rappresentazione tramite istogramma pone numerosi problemi connessi alla determinazione delle classi, sia in rapporto al loro numero (quante classi occorre individuare?) che ai loro estremi (quali valori scegliere per estremi dell'intervallo della classe i-esima, $i = 1, 2, \ldots, k$?). Infatti, cambiando la numerosità delle classi e gli estremi degli intervalli si possono ottenere distribuzioni di frequenze (assolute e relative) piuttosto differenti.
- Il numero k delle classi non deve essere troppo elevato (per evitare che l'istogramma sia troppo dettagliato e senza forma precisa) e nemmeno troppo basso (per evitare di perdere troppe informazioni sulla variabile). Le soluzioni più efficienti richiedono di conoscere preliminarmente alcune misure sintetiche sulla variabile per cui, generalmente, si determina k mediante una serie di tentativi, in funzione del dettaglio che si desidera ottenere.
- Fissato k, tra i vari criteri utilizzati per determinare le ampiezze delle classi, il più diffuso è quello delle *classi equi-ampie*, secondo il quale ciascuna classe ha una ampiezza costante e pari a: $\left(\frac{max(X)-min(X)}{k}\right)$. Il criterio è particolarmente agevole poiché, fissate le modalità estreme ed il numero di classi richieste, la distribuzione di frequenza si completa man mano che arrivano le rilevazioni sulle unità statistiche, sequenzialmente raccolte, e senza necessità di ordinarle (se si conoscono gli estremi).
- Purtroppo, tale criterio cade in difetto quando la variabile è quasi tutta accentrata in uno degli estremi. Oltre che per le variabili di durata (attese agli sportelli, durate di conversazioni telefoniche, sopravvivenza ad una malattia, etc.), questo problema si presenta regolarmente per la distribuzione personale del reddito delle persone fisiche e per numerose altre variabili economiche.
- Inoltre, in presenza di valori eccezionalmente differenti dagli altri, la gran parte delle unità si trova ad essere collocata in qualche classe estrema facendo perdere ogni utile dettaglio alla distribuzione di frequenza. Per esempio, se nella distribuzione dei redditi IRPEF di un comune italiano un solo residente dichiara nell'anno 4 milioni di Euro e si desiderano 10 classi di modalità equi-ampie, si verificherà verosimilmente che quasi tutti i residenti saranno collocati nella prima classe di modalità, perché essa contiene i redditi annui compresi tra 0 e 40 mila Euro.

- Criteri che superano le difficoltà dell'istogramma richiedono strumenti statistici più complessi, e alcuni di essi saranno richiamati nel Capitolo 7.
- A volte, per esigenze pratiche o di calcolo, si ipotizza che la distribuzione di frequenza per variabili continue sia, di fatto, equivalente ad una distribuzione discreta ove i singoli valori sono idealmente i *valori centrali* delle classi di modalità. Questo artifizio permette di sostituire ad una variabile continua una variabile discreta supponendo che ciascuna classe sia ben rappresentata dal suo valore centrale. È evidente che, in molte situazioni reali, tale convenzione può essere molto approssimativa perché all'interno di classi di modalità molto ampie vi può essere un accentramento verso uno dei valori della classe medesima.
- Nelle rappresentazioni grafiche, si possono utilizzare le frequenze assolute oppure quelle relative perché, in entrambi i casi, non si modifica la forma del grafico ma solo la scala delle ordinate. In particolare, se si è interessati a sovrapporre più istogrammi, è necessario utilizzare le frequenze relative.
- Una rappresentazione analoga all'istogramma, molto diffusa in Demografia, è la cosiddetta «piramide delle età» che consiste di due istogrammi accostati che rappresentano la distribuzione di frequenza delle età della popolazione, distintamente per i due sessi (in modo che la coda destra della distribuzione sia verso l'alto e siano affiancate verticalmente l'asse delle ascisse, cioè le classi di età). La forma di tale piramide e la sua evoluzione temporale costituiscono elementi fondamentali per l'analisi e la previsione di importanti fenomeni socio-demografici: natalità, mortalità, nuzialità, previdenza, forza-lavoro, etc.

3.5 La funzione di ripartizione empirica

Se consideriamo la distribuzione delle modalità di una variabile quantitativa X, rilevate sulle n unità della popolazione, *ordinate* dal valore minimo $x_{(1)}$ al valore massimo $x_{(n)}$ e non necessariamente tutte distinte tra loro, e che indicheremo con:

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(i)} \leq \ldots \leq x_{(n)},$$

osserviamo che ciascuna di tali modalità ha frequenza assoluta 1 e frequenza relativa $\frac{1}{n}$.

Si può, quindi, affermare che, nella popolazione:

- la frazione $\frac{1}{n}$ di unità statistiche presenta valori di X che sono inferiori o uguali a $x_{(1)}$;
- la frazione $\frac{2}{n}$ di unità statistiche presenta valori di X che sono inferiori o uguali a $x_{(2)}$; $\ldots$;
- la frazione $\frac{i}{n}$ di unità statistiche presenta valori di X inferiori o uguali a $x_{(i)}$; $\ldots$;
- la frazione $\frac{n}{n} = 1$ di unità statistiche (cioè tutte) presenta valori di X inferiori o uguali a $x_{(n)}$.

La funzione che associa ad ogni valore reale x_0 la frazione delle unità che sono minori o uguali (cioè, non superiori) a x_0 si chiama ***funzione di ripartizione empirica*** calcolata nel valore x_0 (nella letteratura anglosassone, essa viene chiamata *distribution function*).

Pertanto, la *funzione di ripartizione empirica* della variabile X, per ciascun x_0 reale, è definita da:

$$\begin{aligned} F(x_0) &= \text{frequenza relativa delle unità tali che } (X \leq x_0) \\ &= \frac{\text{frequenza assoluta delle unità tali che } (X \leq x_0)}{n} \end{aligned}$$

In pratica, *dopo aver ordinato le modalità di X dal valore minimo al valore massimo*, la funzione di ripartizione è ottenuta «cumulando» progressivamente, al crescere di x, le frequenze relative.

Di conseguenza, $F(x)$ è una funzione definita per ogni x che varia da $-\infty$ a $+\infty$. Essa aumenta di un gradino pari a $\frac{1}{n}$ per ogni valore osservato nella popolazione; così, al variare di x tale funzione cresce di $\frac{1}{n}$ ogni volta che la modalità è presente, e rimane costante quando essa è assente.

Quindi, la funzione di ripartizione è sempre non-decrescente ed è compresa tra 0 e 1, assumendo un andamento spesso simile ad una «esse» allungata (detto, *sigmoide*). Infine, poiché è stata definita per valori *minori o uguali* dell'argomento x, essa è continua da destra, cioè rispetto a valori che stanno alla destra del valore rispetto al quale viene calcolata.

Pertanto, la funzione di ripartizione soddisfa le seguenti proprietà:

$$0 \leq F(x) \leq 1; \qquad F(-\infty) = 0; \qquad F(+\infty) = 1;$$
$$F(x) \text{ è non decrescente}; \quad F(x) \text{ è continua da destra.}$$

Si osservi che la funzione di ripartizione empirica è sempre ben definita, *anche per valori dell'argomento che non sono compresi nel collettivo in esame.*

Infatti, la funzione di ripartizione empirica vale sempre 0 quando x si muove da $-\infty$ sino a che non raggiunge l'estremo inferiore della prima modalità; cresce o non-decresce per i valori di x intermedi a quelli presenti nella popolazione, e raggiunge infine il valore di 1 a partire dal valore massimo sino a $+\infty$.

Esempio 3.5

In una popolazione di 20 unità la variabile X ha manifestato le modalità seguenti (che presentiamo già ordinate, per comodità):

$$\{4, 4, 4, 6, 9, 9, 9, 10, 12, 12, 12, 12, 12, 23, 25, 27, 27, 27, 29, 33\}.$$

Da tali dati si trae la seguente tabella ove, accanto alle frequenze assolute e relative, si ottiene la funzione di ripartizione sommando sequenzialmente le frequenze relative.

Modalità della variabile X	Frequenze assolute	Frequenze relative	Funzione di ripartizione
4	3	0.15	0.15
6	1	0.05	0.20
9	3	0.15	0.35
10	1	0.05	0.40
12	5	0.25	0.65
23	1	0.05	0.70
25	1	0.05	0.75
27	3	0.15	0.90
29	1	0.05	0.95
33	1	0.05	1.00
Totali	20	1.00	===

L'interpretazione risulta immediata: il fatto che la funzione di ripartizione in corrispondenza del valore $X = 9$ sia pari a 0.35 significa che 0.35 (cioè il 35%) è la frazione delle unità statistiche che possiedono la variabile in misura inferiore o uguale a $X = 9$; infatti, tale frazione è ottenuta sommando la quota delle unità che la possiedono in misura $X = 4$ (cioè il 15%) più la quota delle unità che la possiedono in misura $X = 6$ (cioè il 5%) più la quota delle unità che la possiedono in misura $X = 9$ (cioè il 15%).

L'esempio mostra perché, essendo ottenuta per accumulo progressivo, la funzione di ripartizione sia non-decrescente rispetto al suo argomento, come è anche confermato dalla Figura 3.3.

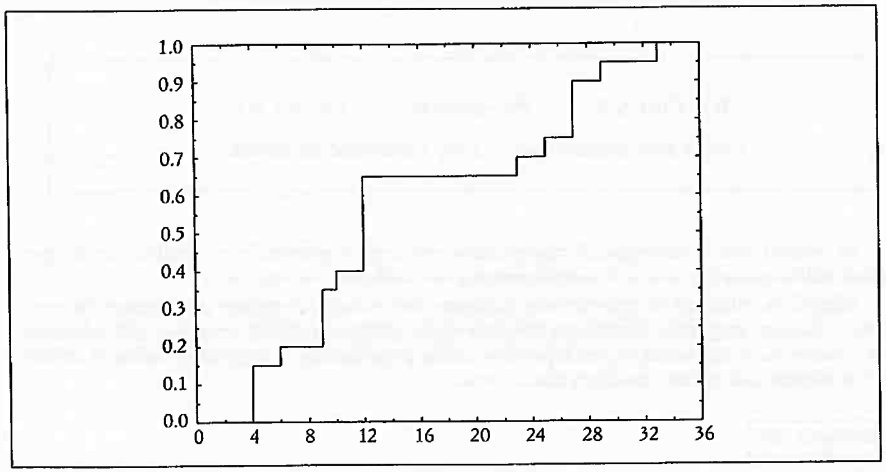

FIGURA 3.4. Funzione di ripartizione per i dati dell'esempio 3.6.

Si osservi che la funzione di ripartizione è ben definita per ogni argomento reale. Così, in questo esempio, anche se la variabile si è manifestata nel collettivo esaminato solo mediante alcuni numeri interi compresi tra 4 e 33, le espressioni $F(2), F(12.43), F(51)$ sono perfettamente lecite. Infatti,

- la quantità $F(2)$ vale 0, perché secondo la definizione della funzione di ripartizione occorre individuare la frazione di unità statistiche che possiedono le modalità in misura minore o uguale a 2, e tale frazione è $\frac{0}{20} = 0$ perché nessuna unità statistica possiede modalità minori o uguali a 2;
- la quantità $F(12.43)$ vale 0.65, perché secondo la definizione della funzione di ripartizione occorre individuare la frazione di unità statistiche che possiedono le modalità in misura minore o uguale a 12.43, e tale frazione è $\frac{13}{20} = 0.65$ perché le unità statistiche con modalità minori o uguali a 12.43 sono quelle che presentano modalità pari a $4, 6, 9, 10, 12$, quindi con frequenze assolute complessivamente pari a 13;
- la quantità $F(51)$ vale 1, perché secondo la definizione della funzione di ripartizione occorre individuare la frazione di unità statistiche che possiedono le modalità in misura minore o uguale a 51, e tale frazione è 1 perché tutte le unità statistiche presentano modalità minori o uguali a 51, essendo la modalità massima uguale a 33.

□

Esempio 3.6

Con i dati di cui all'esempio 3.3, calcoliamo la funzione di ripartizione per la variabile discreta X = «Numero di componenti delle famiglie residenti», in Lombardia e Calabria.

Numero di componenti	Lombardia		Calabria	
	Frequenze relative	Funzione di ripartizione	Frequenze relative	Funzione di ripartizione
1	0.1921	0.1921	0.1700	0.1700
2	0.2411	0.4332	0.2130	0.3830
3	0.2426	0.6758	0.1787	0.5617
4	0.2158	0.8916	0.2077	0.7694
5	0.0759	0.9675	0.1285	0.8979
6	0.0325	1.0000	0.1021	1.0000
Totale	1.0000	======	1.0000	======

L'interpretazione della funzione di ripartizione è immediata: per la Lombardia si ha $F(4) = 0.8916$ mentre per la Calabria $F(4) = 0.7694$. Ciò implica che l'89.16% delle famiglie lombarde sono costituite al più da 4 componenti (cioè da 1, 2, 3 o 4 componenti) mentre il 76.94% delle famiglie calabresi sono costituite al più da 4 componenti.

Ancora, la frazione delle famiglie lombarde e calabresi con più di 5 componenti è ottenuta calcolando: $1 - F(5)$. La frazione delle famiglie con 5 o più componenti è pari a $1 - F(4)$ e, infine, per conoscere il numero di famiglie con componenti compresi tra più di 3 e al massimo 5 (cioè per X appartenente all'intervallo reale: $3 < X \leq 5$) basta calcolare: $F(5) - F(3)$. Si controlla agevolmente che tale frequenza relativa vale 0.2917 per la Lombardia e 0.3362 per la Calabria.

□

Esempio 3.7

La tabella seguente mostra la funzione di ripartizione per la variabile continua X = «Durata dei brani» per il data set COMPACT, di cui all'esempio 3.4 (secondo il criterio delle classi equi-ampie).

Classi di modalità (equi-ampie)	Frequenze relative	Funzione di ripartizione
$30 \leq X \leq 120.182$	0.02685	0.02685
$120.182 < X \leq 210.364$	0.32718	0.35403
$210.364 < X \leq 300.545$	0.36577	0.71980
$300.545 < X \leq 390.727$	0.19799	0.91779
$390.727 < X \leq 480.909$	0.04027	0.95805
$480.909 < X \leq 571.091$	0.02097	0.97903
$571.091 < X \leq 661.273$	0.00923	0.98826
$661.273 < X \leq 751.455$	0.00503	0.99329
$751.455 < X \leq 841.636$	0.00419	0.99748
$841.636 < X \leq 931.818$	0.00168	0.99916
$931.818 < X \leq 1022$	0.00084	1.00000
Totale	1.00000	====

L'ultima colonna, che rappresenta la funzione di ripartizione empirica, è ottenuta sommando sequenzialmente le frequenze relative presenti nella seconda colonna.

L'interpretazione della funzione di ripartizione è immediata. Per esempio, $F(210.364) = 0.35403$ significa che il 35.403% dei brani esaminati ha una durata non superiore a 210.364 secondi. Similmente, il fatto che $F(751.455) = 0.99329$ implica che il 99.329% dei brani ha una durata non superiore a 751.455 secondi e, quindi, appena lo 0.671% (ottenuto da: 100 − 99.329) ha una durata superiore a 751.455 secondi.

Infine, la frequenza relativa dei brani la cui durata è compresa tra 120.182 secondi e 390.727, per esempio, si ottiene dalla ovvia relazione:

$$F(390.727) - F(120.182) = 0.91779 - 0.02685 = 0.89094\ ,$$

cioè i brani la cui durata in secondi è compresa nell'intervallo (120.182, 390.727] costituiscono poco più dell'89% del totale.

□

o Nel campo delle scienze demografiche e sanitarie si utilizza una funzione complementare alla funzione di ripartizione, detta *funzione di sopravvivenza* e definita da: $S(x) = 1 - F(x)$.

In pratica, esaminata una generazione di n individui sopravviventi all'età iniziale (0), si calcolano per ciascun istante successivo (generalmente per ogni anno) il numero di sopravviventi ai tempi $x = 1, 2, \ldots, \omega$ rispetto agli n iniziali, essendo ω l'età massima ipotizzabile per il collettivo.

Il grafico di tale funzione è utile per confrontare la velocità di eliminazione degli individui in circostanze differenti (per esempio, rispetto al sesso, alla regione di residenza, alla dieta, al tipo di professione, all'assunzione di sostanze stupefacenti, al reddito medio percepito, all'inquinamento urbano, etc.).

3.6 Indicatori sintetici delle variabili statistiche

Come abbiamo spesso sottolineato è opportuno (e talora anche necessario) confrontare fenomeni differenti o lo stesso fenomeno in tempi, luoghi, circostanze differenti. Ciò diviene più semplice se il confronto tra le distribuzioni di frequenza avviene mediante misure di sintesi che riassumano importanti aspetti delle variabili in esame. Tali indicatori, evidentemente, devono tener conto sia delle modalità assunte dalle variabili che delle corrispondenti frequenze.

Esempio 3.8

o Supponiamo di voler confrontare il reddito annuo di due cittadini: A possiede un reddito di 10.000 Euro, B possiede un reddito di 30.000 Euro. La situazione espressa da questi due dati statistici (10.000 Euro per A e 30.000 per B), è univoca per cui, fissata la direzione dell'ordinamento, il confronto tra A e B non provoca alcun dubbio e induce a concludere che «A è più povero di B» ovvero che «B è più ricco di A».

o Supponiamo, invece, che le informazioni siano le seguenti: la famiglia di A è composta da tre componenti il cui reddito annuo è rispettivamente 20.000, 18.000, 40.000 mentre la famiglia di B è formata da due componenti il cui reddito annuo è rispettivamente 60.000 e 12.000. In tale situazione, non è univoca una corrispondenza ordinata tra A e B, per molti motivi. Per esempio, il reddito complessivo della famiglia di A (78.000 Euro) è più elevato di quello della famiglia di B (72.000 Euro). Inoltre, B possiede l'elemento più povero fra i cinque (che guadagna 12.000 Euro) ma anche quello più ricco (che guadagna 60.000 Euro); ancora, è evidente che il maggior reddito complessivo di A è la risultante del lavoro di tre unità mentre quello di B viene raggiunto solo con due unità lavoratrici. Infatti, si potrebbe considerare preferibile la situazione della famiglia di B (dicendo che mediamente ciascun componente incassa 36.000 Euro annui) rispetto a quella di A (ove mediamente ciascun componente incassa 26.000 Euro annui).

□

L'esempio ha mostrato come la sintesi vada ricercata in funzione dell'obiettivo da raggiungere: evidenziare il totale guadagnato, esprimere il divario di reddito, e così via.

Naturalmente, la necessaria ricerca degli indicatori non deve mettere in ombra un principio generale che ispira l'intera Statistica, e cioè:

> *Poiché **ogni sintesi costituisce una perdita di informazione** rispetto ai dati elementari, essa va ricercata in modo da minimizzare tale perdita, con riferimento esplicito agli obiettivi della indagine statistica.*

È quindi prioritario esplicitare quali aspetti di una distribuzione di frequenza si intendono esaminare e, successivamente, individuare le misure più idonee. Tale problematica è ben consolidata nella letteratura statistica e concerne lo studio dei seguenti aspetti di una distribuzione di frequenza:

- la ***posizione***, cioè la misura della sua centralità complessiva in rapporto alle modalità e alle rispettive frequenze. La sintesi dovrà essere un valore «rappresentativo» della variabile nella sua globalità, espresso nella stessa unità di misura del fenomeno e capace di «sostituire» in qualche modo tutte le osservazioni;

- la ***variabilità***, cioè la mutevolezza dei dati nella popolazione, ovvero – come spesso si dice, con lo stesso significato – l'attitudine della variabile ad assumere diverse modalità. La sintesi dovrà essere capace di graduare più fenomeni in termini di dispersione, variabilità, cambiamenti, accentramenti rispetto alle osservazioni, stabilità, affidabilità, etc. Come si vede, la variabilità include al suo interno numerosi aspetti che, evidentemente, richiedono misure differenti;

- la *forma*, cioè l'aspetto complessivo della distribuzione di frequenza rispetto a configurazioni standard. In particolare, la sintesi dovrà misurare la simmetria della distribuzione (rispetto ad un punto notevole: per esempio, rispetto ad una misura di posizione), l'accentuazione o l'appiattimento delle modalità più frequenti, il peso relativo degli estremi in rapporto ai valori centrali della distribuzione, e così via.

A loro volta, tutti gli indici statistici di cui ci occuperemo possono essere suddivisi in tre categorie: *assoluti, relativi, normalizzati.*

- Gli ***indici assoluti*** sono misure che possono variare liberamente da un minimo ad un massimo, anche infiniti, ed i cui valori dipendono strettamente dalla natura della variabile che si sta esaminando. Quindi, sono espressi in una unità di misura che dipende dall'unità di misura della variabile in oggetto.

- Gli ***indici relativi*** sono misure svincolate dall'unità di misura perché costituiscono rapporti tra indici assoluti, oppure rapporti tra indici assoluti e loro valori estremi; sono, pertanto «numeri puri» e sono utili per confrontare fenomeni simili, o anche differenti ma logicamente comparabili.

- Gli ***indici normalizzati*** sono particolari indici relativi che variano in un intervallo finito, generalmente in $[0, 1]$ oppure in $[-1, +1]$; possono, quindi, essere utilizzati per effettuare sintesi e confronti tra qualsiasi tipo di fenomeni per i quali essi siano logicamente ed analiticamente calcolabili.

Non tutti gli indici più comuni sono normalizzati, né talvolta è utile normalizzare un indice assoluto: infatti, lo studioso potrebbe essere interessato a valutare gli effetti della sua azione in termini espliciti, riferiti quindi al fenomeno osservato e non ad una sua trasformazione.

Comunque, è sempre possibile normalizzare un indice (o una variabile) di cui sia noto l'insieme dei valori che può assumere (un intervallo finito, un intervallo infinito di valori positivi, l'insieme dei valori reali).

Limitandoci ad esaminare la situazione più frequente di un intervallo finito, supponiamo che l'indice $\mathcal{I}$ sia compreso tra due valori finiti e noti, cioè:

$$\mathcal{I}_{min} \leq \mathcal{I} \leq \mathcal{I}_{max};$$

allora è agevole dimostrare che l'indice $\mathcal{I}^*$ definito da:

$$\mathcal{I}^* = \frac{\mathcal{I} - \mathcal{I}_{min}}{\mathcal{I}_{max} - \mathcal{I}_{min}}$$

è sempre compreso tra 0 e 1, per cui $\mathcal{I}^*$ è un *indice normalizzato.*

Va aggiunto che le *trasformazioni normalizzanti* producono lo stesso effetto (ad esempio, serie di valori compresi tra 0 e 1) anche se sono applicate sulle modalità di qualsiasi variabile quantitativa X, come mostra l'esempio seguente.

Esempio 3.9

Supponiamo di possedere la seguente rilevazione statistica:

{12 *10* 15 18 12 20 25 22 20 24 28 **30** 25 16 20 18 25 22 22 14 18 10 26 22}.

Tale serie si può agevolmente normalizzare tra 0 e 1, secondo la formula precedente, se si sottrae da ogni osservazione il minimo della serie (cioè 10) e si divide il risultato per la differenza tra il massimo ed il minimo dei precedenti valori (cioè per 20, essendo $30 - 10 = 20$).

Procedendo in tal modo si ottiene la seguente serie di modalità normalizzate:

{0.10 0.00 0.25 0.40 0.10 0.50 0.75 0.60 0.50 0.70 0.90 1.00
0.75 0.30 0.50 0.40 0.75 0.60 0.60 0.20 0.40 0.00 0.80 0.60}

Per esempio, i primi valori di tale serie sono stati così ottenuti:

$$0.10 = \frac{12 - 10}{30 - 10},\quad 0.00 = \frac{10 - 10}{30 - 10},\quad 0.25 = \frac{15 - 10}{30 - 10},\quad \ldots$$

□

Ambiente R

Distribuzioni di frequenza

R3.1 Distribuzioni di frequenza e natura delle variabili

Le istruzioni necessarie per ottenere le distribuzioni di frequenza (e le corrispondenti rappresentazioni grafiche) cambiano in funzione della natura della variabile in esame. Infatti, per variabili qualitative e variabili quantitative discrete si pongono in relazione gli elementi di una lista di modalità con le rispettive frequenze ottenute mediante un conteggio di tali modalità. Invece, per variabili continue si pone in corrispondenza un intervallo di valori (cioè una classe, da determinare con qualche criterio) con la frequenza delle unità statistiche che ricadono in quell'intervallo.

Ne consegue che distribuzioni di frequenze per variabili qualitative e quantitative discrete sono univocamente determinate mentre le rappresentazioni grafiche per variabili continue presentano sempre un elemento di soggettività, sono molteplici, e richiedono maggiore cautela nelle interpretazioni.

Una funzione di utilità generale per analizzare variabili (sia se presenti in un vettore che in un *data frame*) è `unique()` che elenca le modalità distinte (quantitative o qualitative) presenti nelle variabili da studiare: è evidente che la distribuzione di frequenza sarà costruita rispetto ad esse. Per esempio, per il data set `studenti`, troviamo:

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> with(data=dati,unique(genere));
[1] 0 1  ### due soli livelli per studenti (0) e studentesse (1)

R==> with(data=dati,unique(diploma));
[1] tec sci alt cla
Levels: alt cla sci tec  ### le 4 modalità qualitative di diploma

R==> with(data=dati,unique(componenti));
 [1]  6  4  8  5  3  1  2  7 12  9 10  ### numerosità dei nuclei
                                       ### familiari

R==> with(data=dati,sort(unique(componenti)));
 [1]  1  2  3  4  5  6  7  8  9 10 12 ### idem, con modalità
                                      ### ordinate
```

Si osservi che, nel data set esaminato, nessuno studente vive in famiglie di 11 componenti. Si ricordi, ancora una volta, che R chiama livelli (=*levels*) le modalità di una variabile qualitativa (che viene chiamata *factor*).

R3.2 Variabili qualitative

Nel caso di variabili qualitative di tipo dicotomico (per esempio, il genere delle unità del data set `studenti`), è sufficiente contare le variabili di tipo logico, `(genere==0)` per gli studenti e `(genere==1)` per le studentesse, e poi sommare mediante la funzione `sum()` i valori in cui tale affermazione logica è vera.

In tal caso, le istruzioni occorrenti ed i risultati corrispondenti sono i seguenti:

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> genere=dati[,2];

R==> studm=sum(genere==0);
R==> freqm=studm/length(genere);percm=100*freqm;
R==> studm;freqm;percm;
[1] 1079
[1] 0.39108
[1] 39.108

R==> studf=sum(genere==1);
R==> freqf=studf/length(genere); percf=100*freqf;
R==> studf;freqf;percf;
[1] 1680
[1] 0.60892
[1] 60.892

R==> studm+studf; freqm+freqf;percm+percf;
[1] 2759
[1] 1
[1] 100
```

Mediante essi si costruisce la Tabella 3.1 che mostra frequenze assolute, relative e percentuali. Le frequenze assolute si potevano anche ottenere con la funzione `table()` nel modo

TABELLA 3.1 *Distribuzione di frequenza della variabile genere.*

	Frequenze		
genere	*assolute*	*relative*	*percentuali*
Maschile	1079	0.39108	39.108
Femminile	1680	0.60892	60.892
Totali	2759	1.00000	100.000

seguente:

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> with(data=dati, table(genere));
genere
   0    1
1079 1680
```

In genere, per variabili qualitative che possiedono più modalità occorre specificare i livelli (*levels*) in modo da utilizzare la funzione `table()` che si applica ad oggetti di tipo

factor (=variabili qualitative). Se, per il data set `studenti`, si considera l'anno dell'indagine come una variabile qualitativa la distribuzione rispetto all'anno delle frequenze (assolute, relative, percentuali, rispettivamente), riportate nella successiva tabella, si ottiene con le istruzioni seguenti:

	Frequenze		
Anno	*assolute*	*relative*	*percentuali*
1988	476	0.17253	17.253
1989	302	0.10946	10.946
1990	303	0.10982	10.982
1991	327	0.11852	11.852
1992	380	0.13773	13.773
1993	361	0.13084	13.084
1994	374	0.13556	13.556
1995	236	0.08554	8.554
Totali	2759	1.00000	100.000

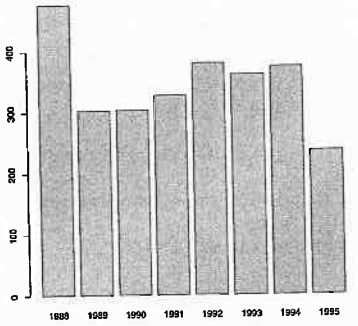

FIGURA R3.1. *Rappresentazione tabellare e grafica della variabile anno.*

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> anno=as.factor(dati[,1]);
R==> table(anno);
anno
1988 1989 1990 1991 1992 1993 1994 1995
 476  302  303  327  380  361  374  236

R==> table(anno)/length(anno);
anno
     1988      1989      1990      1991      1992      1993      1994
1995
0.172526 0.109460 0.109822 0.118521 0.137731 0.130845 0.135556
       0.085538

R==> 100*table(anno)/length(anno);
anno
   1988    1989    1990    1991    1992    1993    1994    1995
17.2526 10.9460 10.9822 11.8521 13.7731 13.0845 13.5556  8.5538
```

Invece, il diagramma a barre della Figura R 3.1 (rappresentazione già discussa nel paragrafo 2.10) si ottiene nel modo seguente:

```
R==> dati=read.table("studenti.txt", header=TRUE);
     ###  in modo esplicito, mediante le seguenti due istruzioni:
R==> anno=as.factor(dati[,1]); plot(anno);
     ###  oppure, direttamente, con una sola istruzione:
R==> with(data=dati, plot(as.factor(anno)));
```

R3.3 Variabili quantitative discrete

Per variabili quantitative discrete, l'elenco delle modalità è una lista di numeri ed R si limita a contare quante unità sono presenti con modalità distinte. Quindi, la distribuzione

di frequenza è rispetto alle modalità presenti nella variabile in esame.

Se le modalità sono interi compresi fra 1 ed m (un numero massimo da specificare), la distribuzione di frequenze (assolute) è ottenuta direttamente tramite la funzione `tabulate()`. L'esempio mostra come ottenere tale risultato per la variabile "numero di componenti" del data set `studenti`. Nello specifico, si osservi che la modalità massima è determinata mediante l'istruzione: `max(componenti)` ed inoltre lo schema finale è ottenuto mediante la funzione `cbind()`, qui applicabile trattandosi di vettori tutti numerici.

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> componenti=dati[,10]; maxcomp=max(componenti);

R==> freq=tabulate(componenti,nbins=maxcomp); n=sum(freq);

R==> freqrel=freq/n; freqperc=100*freqrel;

R==> cbind(1:maxcomp,freq,freqrel,freqperc);

      freq    freqrel  freqperc
 [1,]    1    14 0.00507430  0.507430
 [2,]    2    55 0.01993476  1.993476
 [3,]    3   332 0.12033345 12.033345
 [4,]    4  1057 0.38310982 38.310982
 [5,]    5   874 0.31678144 31.678144
 [6,]    6   304 0.11018485 11.018485
 [7,]    7    91 0.03298296  3.298296
 [8,]    8    24 0.00869880  0.869880
 [9,]    9     3 0.00108735  0.108735
[10,]   10     4 0.00144980  0.144980
[11,]   11     0 0.00000000  0.000000
[12,]   12     1 0.00036245  0.036245
```

Se le modalità discrete sono numeri reali qualsiasi disponibili in serie come $(x_1, x_2, \ldots, x_k)$,

allora la funzione `table()` elenca le modalità discrete uniche in ordine crescente e conta le corrispondenti frequenze assolute:

```
R==> xxx=c(1,1,8,-1,0,0,2,-2,-2,-1,2,8,1,0,-2);
R==> table(xx);
xxx
-2 -1  0  1  2  8
 3  2  3  3  2  2
```

Quindi, le modalità $x_1 = -2$, $x_2 = -1$, $x_3 = 0, \ldots$, presentano frequenze assolute rispettivamente uguali a $n_1 = 3$, $n_2 = 2$, $n_3 = 3, \ldots$

Se in un gruppo i rispondenti forniscono risposte tra 1 e 5 ed in un altro tra 3 e 7, la pedissequa applicazione delle funzioni di `R` produrrebbe distribuzioni di per sé non comparabili perché `R` non elenca le modalità la cui frequenza (conteggio) è pari a 0.

Si considerino, per esempio, le modalità registrate in due gruppi che effettuano valutazioni numeriche:

```
R==> gruppouno=c(1, 1, 5, 5, 3, 2, 4, 3, 2, 2, 4 ,5, 5, 2, 1);
R==> gruppodue=c(3, 3, 5, 5, 3, 7, 7, 6, 4, 4, 4 ,7, 5, 3, 6);
R==> maxmodal=max(gruppouno,gruppodue);
R==> frequno=tabulate(gruppouno,nbins=maxmodal); frequno;
[1] 3 4 2 2 4 0 0
R==> freqdue=tabulate(gruppodue,nbins=maxmodal); freqdue;
```



```
[1] 0 0 4 3 3 2 3
```

In tal modo, avendo calcolato la modalità massima fra i due gruppi, le distribuzioni di frequenze dei due gruppi (di pari numerosità) sono confrontabili perché riferite alle stesse modalità e sono anche rappresentabili "affiancate" in uno stesso diagramma a barre: la Figura `R` 3.2 mostra chiaramente lo spostamento verso valori più elevati delle valutazioni espresse dal secondo gruppo. La rappresentazione è stata ottenuta mediante le seguenti istruzioni (usando qualche trucco grafico: spostare leggermente le ascisse del secondo gruppo, cambiare i toni di grigio delle due rappresentazioni, introdurre una legenda, ...):

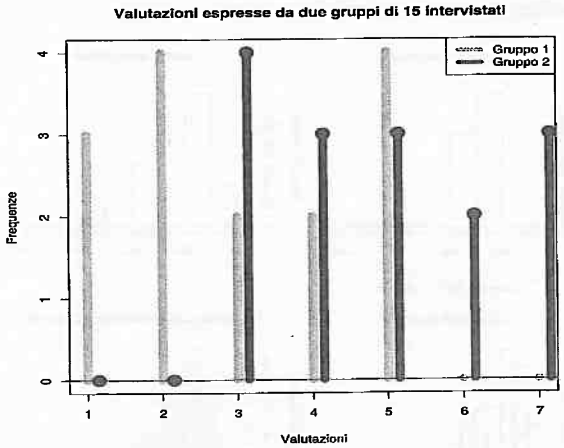

FIGURA R3.2. *Diagrammi a barre "affiancati" con modalità differenti.*

```
R==> plot(1:maxmodal,frequno,type="h",lwd=10,
        xlab="Valutazioni", ylab="Frequenze",
        col=gray(0.7));
R==> points(1:maxmodal,frequno,lwd=10,pch="");
R==> abline(h=0);
R==> shift=0.15;
R==> lines((1:maxmodal)+shift,freqdue,type="h",lwd=10,
        col=gray(0.2));
R==> points((1:maxmodal)+shift,freqdue,lwd=10);
R==> title(main="Valutazioni espresse da due gruppi
        di 15 intervistati");
R==> legend("topright",legend=c("Gruppo 1","Gruppo 2"),
        col=c(gray(0.7),gray(0.2)),
        lwd=c(5,5),lty=1);
```

Con una grafica a colori il risultato è ancora più netto. D'altra parte, esistono in `R` *packages* più raffinati per migliorare queste rappresentazioni.

R3.4 Variabili quantitative continue

Se la variabile è continua, la funzione `hist()` consente di ottenere sia la distribuzione di frequenza dei dati suddivisi in intervalli (con molte opzioni) che la rispettiva rappresentazione grafica mediante *istogramma* (a scelta, delle frequenze assolute oppure relative). Questo è possibile perché la funzione `hist()` genera una lista (*list*) di oggetti alla quale si può attingere in modo semplice. In effetti, se si aggiunge alle opzioni `plot=FALSE` la funzione `hist()` non genera il grafico ma può essere utilizzata per le quantità che essa calcola.

Nel seguito, per il data set `studenti` e la variabile `peso` mostreremo 4 possibili alternative per disegnare istogrammi con differenti opzioni (alcune delle quali discusse nel testo, come quella di Friedman e Diaconis, paragrafo 6.4). Le istruzioni (ed i commenti) dovrebbero essere sufficientemente espliciti.

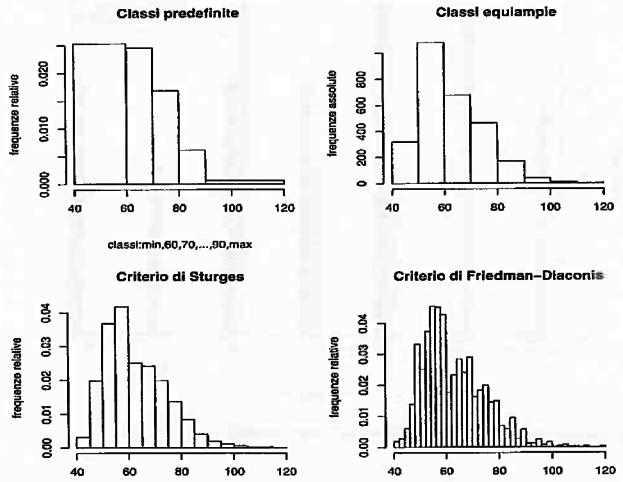

FIGURA R3.3. *Istogramma della variabile peso, in funzione di differenti opzioni.*

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> peso=dati[,4];

R==> par(mfrow=c(2,2)); #suddividi l'area grafica in 4 regioni

R==> ### Grafico 1, opzioni standard, 10 classi equiampie
R==> opz1=hist(peso,freq=FALSE,
           breaks=c(min(peso),60,70,80,90,max(peso)),
           ylab="frequenze relative",main="",xlab="");
R==> title(main="Classi predefinite",
           sub="classi:min,60,70,...,90,max");
R==> freq1=opz1$counts;

R==> # Grafico n.2, freq.assolute, 11 classi equiampie
R==> opz2=hist(peso,freq=TRUE,main="",xlab="",
```



```
           ylab="frequenze assolute",breaks=9);
R==> title(main="Classi equiampie");
R==> freq2=opz2$counts;

R==> # Grafico n.3, freq.relative, STURGES
R==> opz3=hist(peso,freq=FALSE,breaks="sturges",main="",xlab="",
           ylab="frequenze relative");
R==> title(main="Criterio di Sturges");
R==> freq3=opz3$counts;

R==> # Grafico n.4, frequenze relative, FRIEDMAN-DIACONIS
R==> opz4=hist(peso,freq=FALSE,breaks="FD",main="",xlab="",
           ylab="frequenze relative");
R==> title(main="Criterio di Friedman-Diaconis");
R==> freq4=opz4$counts;

R==> par(mfrow=c(1,1)); #ripristina l'area grafica
```

Si noti come l'aspetto grafico dell'istogramma (Figura R 3.3) dipenda molto dal criterio scelto per la numerosità (e la suddivisione) delle classi.

Le istruzioni successive mostrano come ottenere istogrammi per sottoinsiemi della popolazione; nello specifico, mostriamo nella Figura R 3.4 la distribuzione della variabile peso rispetto al genere, sempre per il data set `studenti`. La rappresentazione grafica è ottenuta selezionando la variabile peso che soddisfa la condizione logica su genere (== 0, per gli studenti; == 1 per le studentesse). Poi, mediante l'opzione `breaks=''sturges''`, si disegnano i due istogrammi sulle frequenze relative (aggiungendo l'opzione `freq=FALSE` perché la numerosità dei due sottogruppi è differente ed il confronto tra le frequenze assolute sarebbe ingannevole).

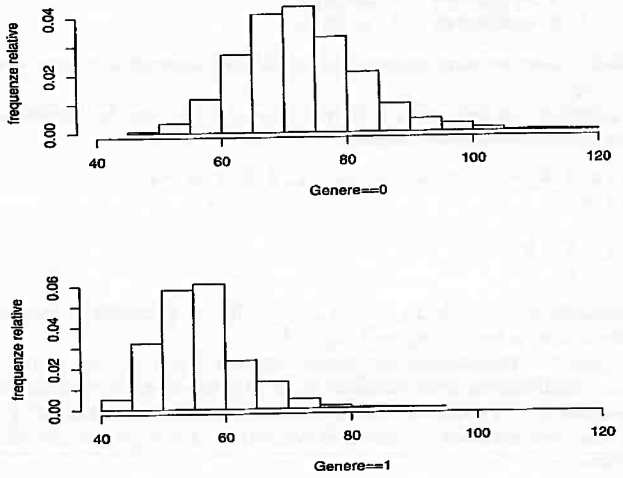

FIGURA R3.4. *Istogramma della variabile peso, in funzione del genere.*

```
R==> dati=read.table("studenti.txt", header=TRUE);
```

```
R==> peso=dati[,4]; genere=dati[,2];
R==> pesom=peso[genere==0]; pesof=peso[genere==1];

R==> par(mfrow=c(2,1))  # suddividi l'area in 2 regioni

R==> # Grafico n.1, freq.relative peso per studenti
R==> hist(pesom,freq=FALSE,breaks="sturges",main="",
          xlab="Genere==0",ylab="frequenze relative",
          xlim=c(40,120));

R==> # Grafico n.2, freq.relative peso per studentesse
R==> hist(pesom,freq=FALSE,breaks="sturges",main="",
          xlab="Genere==1",ylab="frequenze relative",
          xlim=c(40,120));

R==> par(mfrow=c(1,1))  # ripristina l'area del grafico
```

Avendo fissato un intervallo comune per le ascisse (con l'opzione `xlim=c(40,120)`) dei due grafici, ed essendo gli istogrammi riferiti alle frequenze relative (per cui l'area sottesa da entrambi è pari a 1), il loro confronto è corretto: infatti, è ben evidenziato lo spostamento verso modalità più elevate della variabile peso per gli studenti rispetto alle studentesse.

R3.5 Funzione di ripartizione empirica

La funzione di ripartizione empirica si ottiene sommando in modo cumulato le frequenze relative, utilizzando la funzione `cumsum()`. In alternativa, R ha una sua funzione `ecdf()` (=*empirical cumulative distribution function*) la quale opera in modo analogo e che, utilizzata come argomento di `plot()`, produce una rappresentazione stilizzata della funzione di ripartizione empirica.

Per la variabile `componenti` (del nucleo familiare) nel data set `studenti` presentiamo le istruzioni che forniscono il grafico della funzione di ripartizione empirica sia per gli studenti che per le studentesse; ciò viene ottenuto sia tramite `cumsum()` (nella Figura R 3.5, in alto) che tramite `cumsum()` (nella Figura R 3.5, in basso). Nel primo caso, si è usata l'opzione `type=''s''` che unisce i punti mediante uno scalino=*step*; in entrambi i grafici l'ampiezza delle linee agevola la distinzione tra le funzioni. Una legenda completa il grafico.

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> componenti=dati[,10]; genere=dati[,2];
R==> compm=componenti[genere==0]; compf=componenti[genere==1];
R==> maxcomp=max(componenti);

R==> freqm=tabulate(compm,nbins=maxcomp); nm=sum(freqm);
R==> fripm=cumsum(freqm/nm);

R==> freqf=tabulate(compf,nbins=maxcomp); nf=sum(freqf);
R==> fripf=cumsum(freqf/nf);

# Istruzioni per definire l'ambiente grafico di riferimento
par(mfrow=c(2,1)); #suddividi l'area grafica in 2 regioni

### Primo grafico
R==> plot(1:maxcomp,seq(0,1,length=maxcomp),type="n",
```


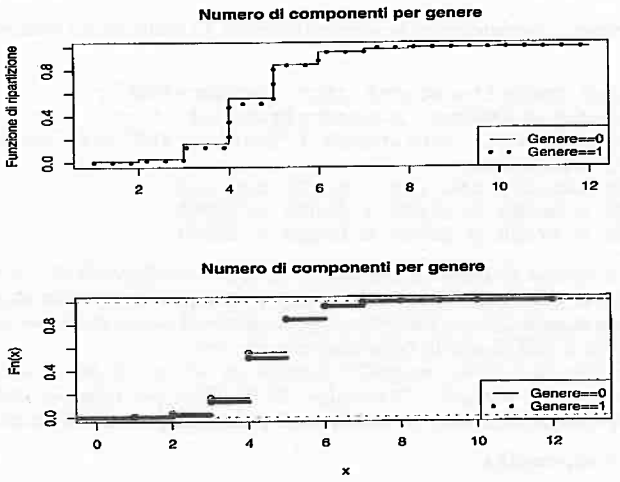

FIGURA R3.5. *Funzione di ripartizione empirica della variabile componenti.*

```
        ylab="Funzione di ripartizione",xlab="",
        xlim=c(1,maxcomp),ylim=c(0,1));
R==> title(main="Numero di componenti per genere");
R==> legend("bottomright",c("Genere==0","Genere==1"),
        lty=c(1,3),lwd=c(1,5));
R==> lines(1:maxcomp,fripm,type="s",
        lty=1,lwd=1); # continua
R==> lines(1:maxcomp,fripf,type="s",
        lty=3,lwd=5); # punteggiata, più consistente

### Secondo grafico
R==> plot(ecdf(compm),main="Numero di componenti per genere");
R==> lines(ecdf(compf),lwd=5);
R==> legend("bottomright",c("Genere==0","Genere==1"),
        lty=c(1,3),lwd=c(1,5));

R==> par(mfrow=c(1,1)); #ripristina l'area grafica
```

Si osservi come, nel caso specifico, la Figura R 3.5 evidenzi una notevole similarità tra le funzioni di ripartizione della variabile `componenti` per i due generi.

R3.6 Normalizzazione di una variabile

Data una variabile quantitativa X (ma potrebbe anche essere un indice statistico costruito su una variabile), per ottenere la variabile normalizzata X^* che assume valori compresi fra 0 e 1 si utilizzano le istruzioni:

```
R==> Xstar=(X-min(X))/(max(X)-min(X))
```

Per esempio, volendo standardizzare le altezze dei primi 15 studenti nel data set studenti, si ha:

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> altezza=dati$altezza; stud10=altezza[1:10];
R==> studstand=(stud10-min(stud10))/(max(stud10)-min(stud10));
R==> stud10; studstand;
 [1] 180 180 173 187 164 186 175 170 180 170
 [1] 0.69565 0.69565 0.39130 1.00000 0.00000
 [6] 0.95652 0.47826 0.26087 0.69565 0.26087
```

Come si vede, al quarto studente (il più alto del gruppo) corrisponde un valore standardizzato pari a 1 e al quinto (il più basso del gruppo) corrisponde un valore standardizzato pari a 0. Il valore standardizzato 0.69565, corrispondente al primo studente, indica che la sua altezza è circa il 79% di quella dello studente più alto.

A tale riguardo, la funzione `range(X)` fornisce un vettore di due elementi di cui il primo è il minimo ed il secondo è il massimo di X. Così, per esempio, sulla variabile `componenti` del dataset `studenti` (esaminata nel paragrafo precedente) si ha:

```
R==> range(componenti)
[1]  1 12
R==> r=range(componenti); r[1]; r[2];
[1] 1
[1] 12
```

Capitolo quarto

Indici statistici di posizione

4.1 Introduzione

Nella ricerca scientifica e tecnologica così come nelle scienze economiche, sociali e politiche, è importante misurare la reale efficacia di interventi sul sistema oggetto di studio. Ciò significa valutare gli effetti complessivi indotti da una causa nota, pur nella mutevolezza ed instabilità dei risultati individuali.

Per questo, occorre una misura di posizione che sintetizzi l'intera distribuzione di frequenza di un certo fenomeno e consenta il passaggio da una pluralità di informazioni (modalità e rispettive frequenze) ad un solo numero, in modo da poter effettuare confronti nel tempo, nello spazio o tra circostanze differenti.

o A tale riguardo, la Statistica ha proposto numerosi indici di posizione per una distribuzione di frequenza e, fra questi, il concetto e la misura di «media» hanno assunto un ruolo centrale al punto che, nell'immaginario collettivo, si ritiene che «la Statistica riguarda le medie». Spesso, l'intera disciplina viene guardata con sospetto allorquando si pretende di sintetizzare le informazioni di una realtà complessa e articolata mediante un unico valore che, si dice, «per rappresentare tutti non rappresenta nessuno».

Famoso è il simpatico sonetto di Trilussa nel quale si sostiene che la Statistica consente di affermare che ciascuno mangia un pollo a testa se metà della popolazione non ne mangia nessuno e l'altra metà ne mangia due. Purtroppo, numerose persone «colte» continuano a citare tale sonetto come se fosse una dimostrazione rigorosa per contestare, a torto o a ragione, l'evidenza dei dati in rapporto a tesi alternative. D'altra parte, i mass-media utilizzano «numeri, percentuali e medie» – raccolti e organizzati in modo variopinto e spettacolare – a sostegno di ipotesi più che come sintesi di fatti. Come vedremo, è un grave errore, concettuale e metodologico, confondere tra la Statistica descrittiva (che mira a fotografare la realtà in modo sintetico ed efficace) e l'Inferenza statistica (che mira ad assumere decisioni e scelte operative sulla base di informazioni limitate e provvisorie), cioè tra momento esplorativo e momento confermativo della realtà.

o Il lettore ha già incontrato, e certamente utilizzato, la media aritmetica nella sua vita quotidiana, nel suo percorso scolastico, nelle sue decisioni operative. Infatti, la media è un concetto primitivo: essa è immanente al comportamento quotidiano anche se, spesso, ciò avviene in modo inconscio.

o Uno dei più antichi riscontri per l'uso della media aritmetica l'abbiamo ritrovato in Geometria quando i babilonesi, per calcolare l'area di un quadrilatero irregolare, emula-

vano il calcolo ben noto dell'area del rettangolo (prodotto dei due lati differenti) facendo il prodotto della semisomma dei lati opposti.

Scopo di questo Capitolo è quello di affrontare, in generale, il problema della migliore sintesi di una distribuzione di frequenza, specificando, poi, in funzione delle regole prefissate, quale misura di volta in volta soddisfi i requisiti richiesti. Emergerà che *la sola media aritmetica non può di per sé esaurire la sintesi di un fenomeno reale* e che, assieme a strumenti grafici elementari, il calcolo congiunto di media e mediana costituisce probabilmente una strategia minimale più efficace.

o Per motivi didattici, la presentazione riguarderà anzitutto il caso di una variabile discreta (sia come serie che come seriazione) evidenziando, poi, le varianti da apportare allo studio di una variabile continua. Successivamente, vedremo in quali casi sia anche possibile pervenire ad indici di posizione per variabili qualitative.

4.2 Il concetto di media

La media è un *concetto* primitivo che viene percepito con immediatezza da ogni persona; tuttavia, la sua *misura* è arbitraria perché il criterio di sintesi che si utilizza dipende strettamente dalle informazioni ritenute più rilevanti e, quindi, dagli obiettivi per i quali la stessa media viene calcolata. A tale riguardo, sono stati introdotti numerosi criteri decisionali, tra i quali discuteremo solo quello di internalità (di Cauchy, 1821) e quello di rappresentatività (di Chisini, 1929).

• Una media secondo Cauchy di una variabile X è qualunque valore reale $\mathbb{M}$ intermedio tra il minimo $min(x)$ ed il massimo $max(x)$ di una distribuzione di frequenza, cioè tale che:

$$min(x) \leq \mathbb{M} \leq max(x).$$

Tale requisito è detto *criterio di internalità* ma – essendo generalmente infiniti i numeri reali che lo soddisfano – esso costituisce più un controllo delle definizioni successive che una soluzione operativa.

• Una media secondo Chisini di una variabile X è quel valore $\mathbb{M}$ intermedio tra il minimo $min(x)$ ed il massimo $max(x)$ il quale, rispetto ad una funzione sintetica delle osservazioni, ne lascia inalterato il valore, cioè tale che:

$$f(x_1, x_2, \ldots, x_n) = f(\mathbb{M}, \mathbb{M}, \ldots, \mathbb{M}).$$

Il senso della definizione deriva dal tentativo di trovare un unico valore $\mathbb{M}$, tale che quando esso viene sostituito a ciascuna modalità di una unità statistica non cambia la predetta funzione.

Ora, perché questo *criterio di rappresentatività* diventi operativo, occorre specificare $f(\ldots)$ in funzione della natura (additiva, moltiplicativa, inversa, etc.) della variabile e della sua trasferibilità. Si possono, quindi, derivare molteplici medie ma, nel seguito, discuteremo solo le più utili e le più diffuse.

4.3 La media aritmetica

Il valore di posizione per eccellenza è la media aritmetica delle osservazioni, spesso definita come «media», senza altre aggettivazioni ed universalmente indicata con il simbolo μ. Essa è ottenuta dal criterio di rappresentatività definendo la funzione $f(x_1, x_2, \ldots, x_n)$ come somma delle modalità $(x_1, x_2, \ldots, x_n)$.

Pertanto, se: $f(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} x_i$, allora tale criterio implica:

$$f(x_1, x_2, \ldots, x_n) = f(\mu, \mu, \ldots, \mu) \Longleftrightarrow \sum_{i=1}^{n} x_i = \mu + \mu + \ldots + \mu$$

e dall'ultima relazione si trae:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Talvolta i dati sono disponibili mediante una seriazione, il che avviene molto spesso quando si fa riferimento a dati pubblicati. In tal caso, si dispone di una distribuzione di frequenza nella quale la modalità x_1 si ripete n_1 volte, la modalità x_2 si ripete n_2 volte, ..., la modalità x_k si ripete n_k volte. Allora, la somma di tutte le modalità (cioè il numeratore della media aritmetica) può essere scritto in modo molto più compatto come: $x_1 n_1 + x_2 n_2 + \ldots + x_k n_k$.

Pertanto, la media aritmetica per variabili discrete – la cui rilevazione è disponibile mediante una distribuzione di frequenza – si può ottenere mediante la formula:

$$\mu = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i} = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_k n_k}{n_1 + n_2 + \ldots + n_k}.$$

Naturalmente, poiché $\frac{n_i}{n} = f_i$, $i = 1, 2, \ldots, k$, quando si dispone delle frequenze relative f_i, per il calcolo della media aritmetica è molto più semplice utilizzare la seguente espressione equivalente:

$$\mu = \sum_{i=1}^{k} x_i f_i = x_1 f_1 + x_2 f_2 + \ldots + x_k f_k.$$

La tabella seguente mostra lo schema di calcolo preferibile per ottenere la media aritmetica (a seconda che si disponga delle frequenze assolute o relative).

Modalità della variabile X	Frequenze		Modalità × frequenze	
	assolute	relative	assolute	relative
x_1	n_1	f_1	$x_1\, n_1$	$x_1\, f_1$
x_2	n_2	f_2	$x_2\, n_2$	$x_2\, f_2$
...	...	...	...	...
x_i	n_i	f_i	$x_i\, n_i$	$x_i\, f_i$
...	...	...	...	...
x_k	n_k	f_k	$x_k\, n_k$	$x_k\, f_k$
Totali	$\sum_{i=1}^{k} n_i = n$	$\sum_{i=1}^{k} f_i = 1$	$\sum_{i=1}^{k} x_i\, n_i$	$\sum_{i=1}^{k} x_i\, f_i = \mu$

o Naturalmente, *le espressioni precedenti non sono formule differenti ma solo tre modi aritmeticamente equivalenti di calcolare la media aritmetica*, la quale è sempre la somma di tutte le modalità di una variabile quantitativa diviso per il numero delle unità statistiche; l'aspetto varia solo perché è differente la disponibilità dei dati.

Esempio 4.1

Supponiamo di aver rilevato su un gruppo di 20 studenti la variabile discreta X = «Misura delle scarpe», ottenendo – per ciascuna delle 20 unità statistiche – le seguenti modalità:

{38, 40, 45, 42, 42, 39, 41, 43, 42, 41, 46, 40, 42, 41, 42, 42, 40, 41,42, 42}.

Trattandosi di dati distinti (=serie), per calcolare la media basta sommare i valori osservati e dividere il totale per la numerosità del collettivo (cioè per 20). Quindi, si ottiene:

$$\mu = \frac{38 + 40 + 45 + \ldots + 41 + 42 + 42}{20} = \frac{831}{20} = 41.55.$$

L'interpretazione del valore così ottenuto è la seguente: ciascuno studente ha per misura di scarpe un numero che non sempre coincide con quello degli altri; volendo scegliere un unico numero che sia rappresentativo delle misure del collettivo di 20 studenti esaminato il valore $\mu = 41.55$ viene scelto perché *la somma di tutte le misure delle scarpe* (cioè 831) è la stessa sia se sostituiamo a ciascuno studente tale valore rappresentativo sia se lasciamo i dati originari nel fare la somma.

Questi stessi dati possono essere organizzati in una distribuzione di frequenza, esplicitando per ciascuna modalità (che varia da 38 a 46, in questo esempio) le frequenze assolute o le frequenze relative, ed operando gli opportuni calcoli come indicati nella tabella seguente:

Modalità della variabile X	Frequenze		Modalità × frequenze	
	assolute	relative	assolute	relative
38	1	0.05	38	1.90
39	1	0.05	39	1.95
40	3	0.15	120	6.00
41	4	0.20	164	8.20
42	8	0.40	336	16.80
43	1	0.05	43	2.15
45	1	0.05	45	2.25
46	1	0.05	46	2.03
Totali	20	1.00	831	41.55

Come si vede, la media aritmetica può essere ottenuta sia utilizzando le frequenze assolute (e, in tale caso, si ha: $\mu = \frac{831}{20} = 41.55$) che le frequenze relative (e, in tale caso, la media è ottenuta direttamente tramite la somma dei prodotti delle modalità moltiplicate per le frequenze relative).

□

Esempio 4.2

Per verificare l'uso della media in problemi di confronto tra due popolazioni, esaminiamo la variabile X = «Misura delle scarpe» per studenti e studentesse. È evidente che trattasi di distribuzioni chiaramente *posizionate* su valori differenti, come mostrano i diagrammi a barre (affiancate) della Figura 4.1: la misura delle scarpe degli studenti è tendenzialmente più elevata della corrispondente misura per le studentesse.

Modalità	Frequenze assolute	
	Maschi	Femmine
34	0	2
35	0	35
36	0	148
37	0	327
38	4	359
39	12	227
40	53	115
41	107	13
42	212	3
43	178	1
44	123	0
45	68	0
46	18	0
47	1	0
48	0	0
49	1	0
Totali	777	1230

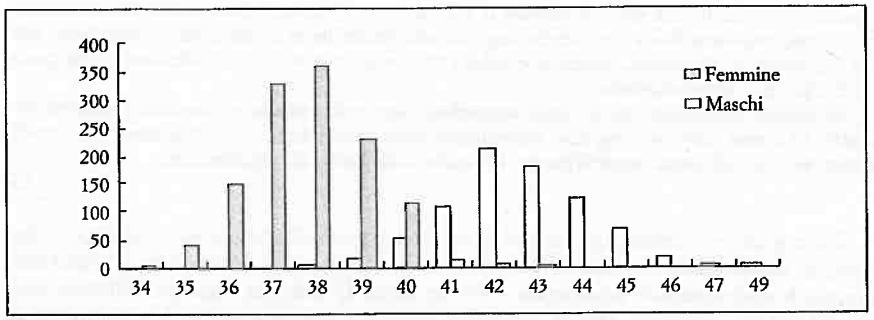

FIGURA 4.1. Diagramma a barre della variabile «misura delle scarpe» (Femmine e Maschi).

Per il calcolo delle corrispondenti medie aritmetiche, si ha:

$$\mu(Maschi) = \frac{38 \times 4 + 39 \times 12 + \ldots + 47 \times 1 + 48 \times 0 + 49 \times 1}{777} = \frac{33081}{777} = 42.575$$

$$\mu(Femmine) = \frac{34 \times 2 + 35 \times 35 + \ldots + 41 \times 13 + 42 \times 3 + 43 \times 1}{1230} = \frac{46517}{1230} = 37.819$$

Le medie sono sintesi delle modalità che si sono registrate sui collettivi in esame e, quindi, il fatto che $\mu(Maschi) > \mu(Femmine)$ non implica che tutti gli studenti possiedano per misura delle scarpe un valore più elevato delle studentesse. Infatti, la media è importante per posizionare tutta intera la distribuzione di frequenza della variabile studiata ma essa, di per sé, non implica alcuna relazione necessaria tra le singole unità statistiche.

□

Infine, quando i dati sono raggruppati in classi (come si verifica per le variabili continue), la soluzione più comune (per quanto approssimata) è quella di utilizzare le espressioni precedenti sostituendo a ciascuna delle classi di modalità $(x_{i-1}, x_i]$ il corrispondente *valore centrale*: $\frac{x_{i-1}+x_i}{2}$, per $i = 1, 2, \ldots, k$.

Esempio 4.3

Calcoliamo la media della variabile «Durata in secondi» dei brani del data set COMPACT (esempio 3.4) che riportiamo nella tabella successiva.

Classi di modalità	Frequenze assolute	Valori centrali della classe	Prodotti dei valori centrali per le frequenze assolute
$0 \vdash - \dashv 60$	5	30	150
$60 - \dashv 120$	27	90	2430
$120 - \dashv 180$	255	150	38250
$180 - \dashv 300$	571	240	137040
$300 - \dashv 600$	312	450	140400
$600 - \dashv 1022$	22	811	17842
Totali	1192	===	336112

Trattandosi di una variabile continua, disponibile per classi di modalità, è necessario utilizzare l'approssimazione di considerare tutti i valori compresi in una certa classe *come se* fossero equivalenti al valore centrale di quella classe. La tabella evidenzia tali valori e presenta nell'ultima colonna i calcoli necessari per addivenire alla determinazione della media aritmetica. Alla fine, si ottiene:

$$\mu \simeq \frac{5 \times 30 + 27 \times 90 + 255 \times 150 + \ldots + 22 \times 811}{1192} = \frac{336112}{1192} = 281.973$$

Quindi, la durata media dei brani viene approssimata mediante il valore di 281.973 secondi, cioè circa 4 minuti e 42 secondi.

Poiché questi dati sono anche disponibili come serie (cioè, nel data set COMPACT sono disponibili le durate dei singoli brani), è possibile calcolare esattamente la media aritmetica. Essa risulta: $\mu = 259.336$, cioè circa 4 minuti e 19 secondi. Come si vede, il calcolo approssimato ha prodotto una sovrastima della media aritmetica, indotta verosimilmente dall'ultima classe della distribuzione di frequenza per la quale l'approssimazione potrebbe essere stata molto grossolana.

□

4.4 Proprietà e difetti della media aritmetica

La diffusione della media aritmetica deriva dalla sua semplicità (concettuale e analitica) ma anche dalle numerose proprietà che soddisfa, alcune delle quali sono *caratterizzanti*, perché valide solo per la media aritmetica. Ci limiteremo, qui di seguito, ad elencare e a commentare solo le principali proprietà:

i) la media aritmetica è sempre compresa tra il minimo e il massimo delle modalità della variabile;

ii) se definiamo *scarti dalla media* (oppure, semplicemente, *scarti*) le quantità: $(x_i - \mu)$ per ogni $i = 1, 2, \ldots, n$, allora la somma degli scarti dalla media aritmetica è nulla.

Infatti, essendo $\sum_{i=1}^{n} x_i = n\mu$, si dimostra agevolmente che:

$$\sum_{i=1}^{n} (x_i - \mu) = \sum_{i=1}^{n} x_i - n\mu = n\mu - n\mu = 0.$$

Questo risultato è di fondamentale importanza perché mostra che la media costituisce il *baricentro* di una distribuzione di frequenza. Infatti, essendo la media μ interna ai valori delle modalità di X, alcuni scarti saranno positivi, alcuni eventualmente nulli, ed altri negativi. Ebbene, questa proprietà afferma che scarti positivi e negativi (individualmente presi, nel caso delle serie, o moltiplicati per le frequenze, se i dati sono disponibili come seriazione) si compensano esattamente.

iii) se la variabile X ha media aritmetica μ, allora la variabile $\alpha X+\beta$ (ottenuta da X mediante una combinazione lineare con coefficienti costanti α e β) possiede media aritmetica pari a: $\alpha\mu+\beta$.

Infatti, se X presenta modalità $(x_1, x_2, \ldots, x_n)$, allora la variabile $\alpha X+\beta$ presenterà modalità $(\alpha x_1+\beta, \alpha x_2+\beta, \ldots, \alpha x_n+\beta)$, per cui la media di quest'ultima sarà:

$$\frac{1}{n}\sum_{i=1}^{n}(\alpha x_i+\beta) = \alpha\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)+\frac{1}{n}\sum_{i=1}^{n}\beta = \alpha\mu+\beta.$$

Tale proprietà è detta *linearità* della media aritmetica ed implica, in particolare:

a) se si aggiunge o si sottrae una costante alla variabile X, la rispettiva media sarà modificata dello stesso ammontare (basta porre $\alpha = 1$ nella dimostrazione);

b) se la variabile X è moltiplicata per un coefficiente costante, anche la media ne risulterà moltiplicata per lo stesso ammontare (basta porre $\beta = 0$ nella dimostrazione).

iv) la media aritmetica è l'unico valore che rende minima la somma degli scarti al quadrato.

Per una certa rilevazione statistica, la quantità $f(c) = \sum_{i=1}^{n}(x_i-c)^2$ dipende ovviamente dal valore di c; ebbene, tale proprietà afferma che $f(c)$ sarà minima se e solo se c coincide con μ, cioè se gli scarti al quadrato sono calcolati rispetto alla media aritmetica.

Questa proprietà è caratterizzante per la media aritmetica e produce importanti conseguenze statistiche.

Le numerose proprietà della media aritmetica non possono nascondere il fatto che essa resta un indicatore sintetico e come tale può essere impropriamente utilizzata a fini comparativi quando l'aggregazione nasconde aspetti differenziali di un fenomeno. I due esempi successivi mostrano alcuni dei rischi derivanti da una interpretazione disinvolta e scientificamente errata della media.

Esempio 4.4

Per il data set STUDENTI, calcolando la media aritmetica per la variabile Peso (in kg) e per la variabile Altezza (in cm) sia per la popolazione dei fumatori che per quella dei non fumatori, otteniamo i seguenti risultati:

Popolazione	Numerosità	Altezza media	Peso medio
Fumatori	703	170.27454	65.24780
Non Fumatori	2056	168.58025	62.74212

Poiché risulta che gli studenti fumatori sono mediamente più alti di quasi 2 cm e più robusti mediamente di quasi 2.5 kg rispetto agli studenti non fumatori, si potrebbe arguire, a torto, che il fumo fa bene! In realtà, nel data set STUDENTI, il 46% dei fumatori è costituito da maschi mentre il 73% dei non fumatori è costituito da femmine. La diversa struttura fisica dei maschi rispetto alle femmine fa sì che nella popolazione dei non fumatori la prevalenza femminile riduca notevolmente le medie aritmetiche del peso e dell'altezza a loro associate.

In questa situazione (ed in altre analoghe) l'uso della media aritmetica è improprio. Infatti, vi è una variabile che non viene esplicitata nelle analisi (qui, è il sesso dell'unità statistica) che influenza sensibilmente la media delle variabili considerate. □

La media aritmetica ha un grave difetto (particolarmente serio per popolazioni di numerosità modesta) nel caso in cui delle unità statistiche presentano modalità notevolmente diverse da tutte le altre. L'esempio seguente illustra una tale situazione estrema, che si potrebbe verificare sia quando vi è una diversità reale nella popolazione che quando la diversità è artificiale, perché è stata indotta da un errore di rilevazione.

Esempio 4.5

Supponiamo che in due condomini A e B il reddito familiare mensile delle 5 famiglie (=unità statistiche) che vi abitano, espresso in migliaia di Euro, sia sintetizzato dalla tabella seguente:

Condominio	Modalità rilevate sulle unità statistiche	Media aritmetica
A	{ 1, 2, 3, 4, 5}	3
B	{ 1, 2, 3, 4, 50}	12

Ora, mentre in A la media $\mu = 3$ può essere utilizzata come una sintesi di valori compresi tra 1 e 5, in B la presenza di un solo valore atipico (=50) ha completamente distrutto la capacità rappresentativa della media aritmetica. Infatti, affermare che $\mu = 12$ nel condominio B non produce una misura di posizione adeguata né per i primi 4 condomini né per l'ultimo. □

Questo difetto della media aritmetica deriva dalla sua proprietà di essere baricentro della distribuzione: infatti, un valore eccezionalmente diverso da tutti gli altri attrae il baricentro nella sua direzione, sino a produrre per μ un valore numerico che è solo un mero punto di equilibrio per la somma delle modalità, e non più un valore equamente rappresentativo di tutti gli altri. Si intuisce da questa situazione che una misura di rappresentatività della media deve essere connessa alla dispersione delle modalità, come sarà evidente nel Capitolo seguente.

Il rischio derivante dalla presenza di dati anomali o errati si è accresciuto nei tempi recenti per l'uso massiccio ed automatico di masse di dati che pochi controllano accuratamente prima di effettuare elaborazioni statistiche.

Il fatto che *anche una sola modalità* completamente diversa da tutte le altre possa rendere la media non più rappresentativa del complesso della posizione della variabile mostra che tale indice di posizione ha una scarsa *resistenza* (robustezza) in presenza di valori eccezionali. Vedremo, alla fine del paragrafo 4.8, come limitare i danni derivanti da tale inconveniente.

4.5 La media aritmetica ponderata

Nella definizione di media aritmetica, le modalità (e quindi le unità statistiche) concorrono «alla pari» nella determinazione della media: infatti, ciascuna modalità «conta» per $\frac{1}{n}$ nella somma dei prodotti che definisce la media μ.

Tuttavia, esistono numerose situazioni reali nelle quali tale approccio non è corretto perché le unità statistiche possiedono importanza, esperienza o credibilità differenti tra loro.

- In una competizione sportiva la competenza dei giudici cresce con la loro esperienza e, quindi, potrebbe essere opportuno ponderare ciascun voto espresso da un giudice con il numero di gare internazionali cui ha partecipato. In tal modo, nella valutazione media complessiva, conteranno di più i voti espressi dai giudici con maggiore esperienza.
- Nel sintetizzare le opinioni circa le aspettative della produzione industriale in Italia, è ragionevole dare maggiore rilevanza alle opinioni espresse da chi è più pesantemente coinvolto in tale produzione, il che avviene ponderando tali opinioni con una misura di tale rilevanza (per esempio, con il numero di addetti, con il fatturato annuo, etc.).
- Un modo simile di ponderazione sono la latitudine media (cioè la media aritmetica delle latitudini delle unità statistiche ponderate con le modalità del fenomeno) e la longitudine media (cioè la longitudine delle unità statistiche ponderata con le modalità del fenomeno) che vengono ottenute mediante medie ponderate: esse determinano sul territorio una sorta di «centro medio» su cui pensare idealmente addensato il fenomeno in esame. Il confronto tra tali centri medi, rispetto a fenomeni simili o nel tempo, costituisce una interessante rappresentazione grafica. Pubblicazioni ISTAT ed alcuni atlanti geografici contengono, spesso, tali rappresentazioni.
- Dopo la riforma universitaria, che ha introdotto il sistema dei «crediti» per misurare l'impegno richiesto per ciascun insegnamento universitario, è giusto valutare la carriera universitaria compiuta da uno studente ponderando i singoli voti con i crediti acquisiti nel corrispondente esame. In tal modo, si fornisce maggiore rilevanza ai risultati conseguiti nelle materie che hanno richiesto un maggior impegno di tempo.

Ora, la media aritmetica è stata definita come la somma dei prodotti tra le modalità x_i ed un peso *costante* (cioè 1) e dividendo poi per la somma di tali pesi (cioè per n). In modo analogo, si può definire la media aritmetica ponderata come la somma dei prodotti delle modalità x_i per dei pesi *variabili* w_i (che crescono con l'importanza che si intende attribuire alla modalità x_i) e dividendo poi per la somma di tali pesi w_i.

Ciò avviene agevolmente attribuendo un peso non-negativo w_i a ciascuna modalità e definendo la ***media aritmetica ponderata*** μ_w nel modo seguente:

$$\mu_w = \frac{\sum_{i=1}^{k} x_i\, w_i}{\sum_{i=1}^{k} w_i} = \frac{x_1\, w_1 + x_2\, w_2 + \ldots + x_k\, w_k}{w_1 + w_2 + \ldots + w_k} .$$

È agevole dimostrare che anche la media aritmetica ponderata è il baricentro delle osservazioni (cui vengono applicati i pesi w_i non più costanti) e che per essa valgono le stesse proprietà della media, discusse nel paragrafo 4.4.

Esempio 4.6

Per verificare le conseguenze dell'uso della media ponderata rispetto alla media aritmetica, presentiamo il caso di due studenti (che indichiamo con X e Y), i quali hanno entrambi concluso il loro iter formativo dei 18 esami previsti per un corso di Laurea triennale e che hanno anche ricevuto gli stessi voti (ma ad esami differenti). L'esempio presuppone che, dei 180 previsti nel triennio, 140 crediti sono conseguiti mediante esami e 40 crediti mediante altre attività formative.

Per entrambe le distribuzioni presentiamo sia il calcolo della media aritmetica μ (che evidentemente coincide per i due studenti) che della media aritmetica ponderata μ_w che tiene conto del differente numero dei crediti di ciascun esame.

Esami	Studente X			Studente Y		
	Voti (x_i)	Crediti (w_i)	Voti×Crediti ($x_i\, w_i$)	Voti (y_i)	Crediti (w_i)	Voti×Crediti ($y_i\, w_i$)
1	25	4	100	25	4	100
2	30	6	180	30	6	180
3	30	4	120	27	4	108
4	28	5	140	24	5	100
5	22	12	264	22	12	264
6	27	8	216	30	8	240
7	25	9	225	25	9	225
8	30	9	270	28	9	252
9	24	7	168	30	7	210
10	30	5	150	30	5	150
11	20	10	200	27	10	270
12	27	5	135	20	5	100
13	26	10	260	28	10	280
14	28	6	168	26	6	168
15	22	10	220	30	10	300
16	30	4	120	22	4	88
17	30	4	120	22	4	88
18	22	12	264	30	12	360
Totali	476	130	3320	476	130	3443

Pertanto, si ottiene:

$$\begin{aligned}
\mu_X &= \frac{25+30+30+\ldots+22}{18} = \frac{476}{18} = 26.444 \\
\mu_{wX} &= \frac{25\times 4+30\times 6+30\times 4+\ldots 22\times 12}{4+6+\ldots+12} = \frac{3320}{130} = 25.538 \\
\mu_Y &= \frac{25+30+27+\ldots+30}{18} = \frac{476}{18} = 26.444 \\
\mu_{wY} &= \frac{25\times 4+30\times 6+27\times 4+\ldots 30\times 12}{4+6+\ldots+12} = \frac{3443}{130} = 26.485
\end{aligned}$$

Come si vede, a parità di media aritmetica, la ponderazione con i crediti provoca una sensibile differenza nella valutazione media della preparazione universitaria complessivamente conseguita da X e Y. Il divario è causato dalla circostanza per cui i voti più elevati di X sono stati conseguiti in esami cui sono attribuiti pochi crediti, mentre per Y vale esattamente il contrario.

□

Un'applicazione importante della media aritmetica ponderata, con riflessi notevoli per il dibattito economico, si ha con riferimento ai «*numeri indice*» (detti anche «numeri indici»), di cui discuteremo qui solo gli elementi essenziali. Per ogni utile approfondimento è necessario riferirsi ai testi di Statistica Economica e ai manuali metodologici e alle note tecniche dell'ISTAT.

o Sul sito dell'ISTAT (*www.istat.it*) sono presenti definizioni, metodi di calcoli, riferimenti normativi e dati periodicamente aggiornati sui principali numeri indice calcolati in Italia (dei prezzi, dei salari, della produzione, del commercio internazionale, del costo di un fabbricato residenziale, etc.) assieme ad un loro adeguato confronto internazionale. Inoltre, altri numeri indice – soprattutto nel mercato finanziario – vengono quotidianamente seguiti dagli operatori di borsa, registrandone le variazioni addirittura in tempo reale: MIBTEL, Dow Jones, Standard&Poors500, Nasdaq, etc.

I ***numeri indice*** sono ottenuti rapportando due valori differenti di uno stesso fenomeno in circostanze differenti. Essi sono utili per confrontare l'ammontare di un fenomeno in tempi e/o in luoghi differenti mediante una misura che è sempre positiva e indipendente dall'unità di misura: è prassi moltiplicare il rapporto per 100, per facilitare il confronto delle variazioni in termini percentuali.

o Così, un numero indice dei prezzi pari a 100 indica assenza di variazioni, un numero indice di 90 implica una diminuzione nei prezzi del 10%, un numero indice di 104 implica un incremento dei prezzi del 4%. Nei tre casi, il complemento a 100 esprime la differenza relativa (percentuale) registrata tra le due occasioni a confronto.

o I numeri indice più diffusi riguardano le *variazioni dei prezzi* che consentono, nel tempo e sul territorio, di valutare la dinamica dell'inflazione mediante confronti tra i prezzi (rispetto ad un anno-base ovvero rispetto ad una media nazionale, per esempio). Si parla di *inflazione congiunturale* quando la variazione dei prezzi è confrontata rispetto al periodo immediatamente precedente (generalmente, il mese precedente); invece, si parla di *inflazione tendenziale* quando la variazione dei prezzi è confrontata rispetto allo stesso periodo dell'anno precedente (generalmente, lo stesso mese dell'anno precedente).

Ora, il problema del confronto tra prezzi deriva dal fatto che occorre tenere conto di più beni e servizi tra loro molto eterogenei: un modo semplice è quello di confrontare il valore monetario del tempo t con l'analogo valore del tempo 0. Il valore monetario di un complesso di beni e servizi si ottiene moltiplicando i loro prezzi per le quantità commercializzate (vendute, acquistate, prenotate, etc.) sia nel tempo t che nel tempo 0. Ebbene, il numero indice dei prezzi è il rapporto (generalmente, moltiplicato per 100) tra il valore monetario al tempo t rispetto all'analogo valore calcolato al tempo 0.

Il problema metodologico rilevante è quello di definire qual è la quantità di riferimento al variare del tempo: se essa è *fissa*, l'indice sintetico cui si perviene si

chiama di Laspeyres; se essa è *variabile*, l'indice sintetico corrispondente si chiama di Paasche.

Supponiamo, allora, di voler confrontare i prezzi di m beni o servizi in due tempi differenti: indichiamo con 0 il tempo di riferimento (detto anche *tempo-base*) e con t il tempo di calcolo, per $t = 1, 2, \ldots$ Oltre ai prezzi rilevati sui beni o servizi, supponiamo di conoscere anche le corrispondenti quantità (scambiate, richieste, etc.), come nella tabella seguente:

Tempi	Prezzi	Quantità	Valori (con q_{i0})	Valori (con q_{it})
0	$p_{10}, p_{20}, \ldots, p_{m0}$	$q_{10}, q_{20}, \ldots, q_{m0}$	$\sum_{i=1}^{m} p_{i0} q_{i0}$	$\sum_{i=1}^{m} p_{i0} q_{it}$
t	$p_{1t}, p_{2t}, \ldots, p_{mt}$	$q_{1t}, q_{2t}, \ldots, q_{mt}$	$\sum_{i=1}^{m} p_{it} q_{i0}$	$\sum_{i=1}^{m} p_{it} q_{it}$

Mediante tali informazioni si possono definire differenti numeri indice, tra i quali i due più importanti sono i seguenti:

- L'***indice di Laspeyres*** calcolato in ciascun tempo t rispetto al tempo-base 0 è ottenuto rapportando il valore monetario dei beni e servizi di ciascun tempo t con il valore monetario del tempo 0, essendo tali valori tutti calcolati con le *quantità fisse* del tempo 0. Pertanto, esso è definito mediante la formula seguente:

$$I_{0,t}^{L} = \frac{\sum_{i=1}^{m} p_{it}\, q_{i0}}{\sum_{i=1}^{m} p_{i0}\, q_{i0}} = \frac{\frac{\sum_{i=1}^{m} p_{it}\, q_{i0}}{\sum_{i=1}^{m} q_{i0}}}{\frac{\sum_{i=1}^{m} p_{i0}\, q_{i0}}{\sum_{i=1}^{m} q_{i0}}}.$$

L'ultima espressione mostra che l'indice di Laspeyres è null'altro che il rapporto tra la media ponderata dei prezzi al tempo t e la media ponderata dei prezzi al tempo 0, ove per entrambe i pesi w_i sono ora le quantità q_{io} del tempo 0.

o L'indice di Laspeyres è quello più utilizzato sia per la semplicità di calcolo che per la immediatezza della sua interpretazione. Infatti, esso confronta – al variare del tempo t – il valore monetario di un *paniere fisso* di beni e servizi (perché le quantità sono quelle definite al tempo-base 0) rispetto all'analogo valore di quel paniere calcolato nel tempo-base 0.

Peraltro, il suo calcolo è molto semplice perché (una volta fissate le quantità q_{i0} presenti nel paniere al tempo-base 0) l'indice di Laspeyres richiede di conoscere periodicamente *solo* i prezzi p_{it}, al variare del tempo t di calcolo.

Quest'ultima considerazione diviene immediata se si osserva che l'indice di Laspeyres si può scrivere nel modo seguente:

$$I^L_{0,t} = \sum_{i=1}^{m} p_{it} \quad \frac{q_{i0}}{\sum_{i=1}^{m} p_{i0}\, q_{i0}} = \sum_{i=1}^{m} p_{it}\, w_{i0},$$

dove i pesi: $w_{i0} = \frac{q_{i0}}{\sum_{i=1}^{m} p_{i0}\, q_{i0}}$, che misurano l'importanza relativa del bene i-esimo all'interno del paniere di beni e servizi, sono evidentemente costanti al variare di t.

o La quasi totalità dei numeri indice dei prezzi calcolati in Italia e in Europa utilizzano la formula di Laspeyres.

- L'***indice di Paasche*** calcolato in ciascun tempo t rispetto al tempo-base 0 è ottenuto rapportando il valore monetario dei beni e servizi di ciascun tempo t con il valore monetario al tempo 0, essendo tali valori calcolati con le *quantità variabili* del tempo t. Pertanto, esso è definito mediante la formula seguente:

$$I^P_{0,t} = \frac{\sum_{i=1}^{m} p_{it}\, q_{it}}{\sum_{i=1}^{m} p_{i0}\, q_{it}} = \frac{\frac{\sum_{i=1}^{m} p_{it}\, q_{it}}{\sum_{i=1}^{m} q_{it}}}{\frac{\sum_{i=1}^{m} p_{i0}\, q_{it}}{\sum_{i=1}^{m} q_{it}}}$$

L'ultima espressione mostra che l'indice di Paasche è il rapporto tra la media ponderata dei prezzi al tempo t e la media ponderata dei prezzi al tempo 0, ove per entrambe i pesi w_i sono ora le quantità q_{it} del tempo t.

o Da ciò si deduce che, generalmente, il calcolo dell'indice di Paasche è più laborioso dell'indice di Laspeyres perché richiede, *per ciascun tempo t*, la conoscenza non solo dei prezzi p_{it} ma anche delle quantità q_{it}. Esso è di agevole calcolo, invece, in tutte quelle situazioni in cui vengono contestualmente elaborati sia i prezzi che le quantità, come per esempio nelle contrattazioni dei titoli in Borsa.

Esempio 4.7

Supponiamo, per semplicità, che i beni e servizi da confrontare siano solo 5 e che si vogliano costruire gli indici di Laspeyres e di Paasche del tempo t rispetto al tempo-base 0, sulla scorta delle informazioni riassunte nella tabella seguente:

Tempi	Prezzi (in Euro)	Quantità (in unità di commercio)
0	10, 200, 1000, 50, 300	200, 100, 2, 100, 250
t	12, 220, 900, 50, 450	220, 110, 4, 108, 200

Mediante tali dati, i due indici sono così calcolati:

$$I^L_{0,t} = \frac{12\times 200 + 220\times 100 + 900\times 2 + 50\times 100 + 450\times 250}{10\times 200 + 200\times 100 + 1000\times 2 + 50\times 100 + 300\times 250} = \frac{143700}{104000} = 1.38173$$

$$I^P_{0,t} = \frac{12\times 220 + 220\times 110 + 900\times 4 + 50\times 108 + 450\times 200}{10\times 220 + 200\times 110 + 1000\times 4 + 50\times 108 + 300\times 200} = \frac{125840}{93600} = 1.34444$$

Quindi, l'indice di Laspeyres segnala tra i due periodi un incremento medio dei prezzi del 38% mentre l'indice di Paasche segnala un incremento medio del 34%.

Si noti la differente interpretazione dei due indici: mentre l'indice di Laspeyres confronta il valore del *paniere prefissato al tempo-base* di 143700 Euro del tempo t con il valore di 104000 Euro del tempo 0, l'indice di Paasche confronta il valore del *paniere determinato dalle quantità variabili* del tempo t di 125840 Euro del tempo t con il valore di 93600 Euro del tempo 0. □

I numeri indice dei prezzi sono regolarmente utilizzati nella vita economica e sociale, e sono diffusi e commentati periodicamente nei mezzi di comunicazione; per questo ci sembra importante puntualizzare alcuni aspetti di natura statistica al fine di una loro corretta utilizzazione.

o L'indice dei prezzi misura e sintetizza le variazioni dei prezzi nel tempo. Sulla base degli indici, è quindi impossibile fare confronti relativi al livello dei prezzi di due prodotti diversi o di due città diverse. Tra le affermazioni inesatte che spesso compaiono nei mass-media va segnalata quella riguardante la determinazione della «città più cara» in un certo mese, ottenuta confrontando le rispettive misure di inflazione mensile.

Come opportunamente avverte l'ISTAT, il tasso di variazione fa riferimento al periodo indicato per cui, per esempio, la constatazione che Milano registra in Febbraio 2004 un tasso di variazione di +0.35% mentre Bologna registra +0.21% implica solo che tra Febbraio e Gennaio 2004 la velocità dei prezzi è stata più elevata a Milano rispetto a Bologna; ma Bologna poteva, comunque, possedere un livello generale dei prezzi di partenza notevolmente più elevato di Milano. In breve, i tassi di variazione misurano la velocità e non il livello dei prezzi dei collettivi che si confrontano.

o L'indice dei prezzi non misura la variazione della spesa delle famiglie. Quest'ultima, infatti, risente delle modificazioni sia del reddito disponibile sia dei comportamenti di consumo, indotti eventualmente anche dalle variazioni dei prezzi.

o La rilevazione dei prezzi al consumo è fatta su un paniere di prodotti mantenuto costante solo per un anno e composto dai prodotti e dai servizi di maggiore fruizione da parte dei consumatori.

o La rilevazione ha per oggetto prezzi effettivamente praticati e non prezzi di listino o tariffe annunciate (per esempio, il prezzo praticato direttamente dall'esercente, dai rivenditori, etc.). Il prezzo rilevato è quello per il pagamento pieno, rilevato al netto degli sconti eventualmente praticati dai rivenditori. Il prezzo scontato è rilevato anche per i saldi e per le campagne promozionali, allorché esse ricorrano con sufficiente regolarità. Attualmente, i punti vendita presso cui vengono rilevati i prezzi dei prodotti contenuti nel paniere sono circa 13.000.

4.6 Moda e mediana

In questo paragrafo ed in quello successivo introdurremo altri indici di posizione che, in vario modo, sostituiscono o completano il significato e la misura della media aritmetica.

■ La *moda* Mo di una distribuzione di frequenza è la modalità cui corrisponde la massima frequenza, assoluta o relativa.

o Sintetizzare una variabile X tramite la sua moda significa assumere come valore «più rappresentativo» della distribuzione quello che si è verificato più spesso di tutti gli altri. Il che implica che se si sceglie tra più preferenze il valore modale si crea il «minor numero di scontenti», perché la moda è la modalità della variabile espressa dal gruppo di maggioranza relativa.

o In popolazioni di ammontare limitato può verificarsi che più modalità presentino la stessa frequenza massima e, in tal caso, si parla di distribuzioni bi-modali o multi-modali. Ciò deriva dalla presenza di sotto-popolazioni che per ragioni strutturali, per errore o per comodità vengono esaminate congiuntamente.

Evidentemente, l'uso della moda è di qualche rilevanza solo nel caso che essa sia unica e con una frequenza ben differenziata rispetto a quella delle altre modalità.

D'altra parte, la moda è un indice di posizione di immediata determinazione e ben interpretabile nei termini del problema perché, a differenza delle medie analitiche (cioè quelle ottenute mediante un calcolo algebrico), la moda è sicuramente il valore di una modalità tra quelle osservate.

Infine, per sua natura la moda – a differenza della media aritmetica – si può determinare *anche per variabili qualitative.*

o D'altra parte, le obiezioni all'uso della moda sono numerose e, talora, fondate. Se, per esempio, due o più modalità – anche distanti tra loro – presentano frequenze simili, la determinazione di una fra loro può dipendere solo da qualche osservazione isolata e non è, quindi, «rappresentativa» della popolazione in esame.

o Inoltre, soprattutto per variabili discrete rilevate su un gran numero di unità statistiche, può avvenire che essa non sia efficace per discriminare situazioni differenti tra loro. Per esempio, la moda del numero di componenti per famiglia in molte regioni d'Italia è la stessa, anche se indubbiamente vi è una posizione differente tra la distribuzione dei componenti delle famiglie nelle varie regioni.

Esempio 4.8

o Per la variabile (qualitativa) «Diploma secondario» di provenienza degli studenti di cui all'esempio 3.1 si deriva immediatamente che la moda è la modalità (=attributo) «Tecnico» essendo questa la modalità più numerosa tra le quattro prese in considerazione.

o Similmente, esaminando i dati di cui all'esempio 3.2 per la variabile «Giorno preferito della settimana per andare al cinema» si deduce subito che la moda è il Sabato essendo questi il giorno maggiormente prescelto.

o La popolazione di 20 unità statistiche:

$$\{1, 2, 2, 2, 2, 3, 4, 6, 6, 7, 7, 7, 8, 8, 9, 9, 9, 10, 11, 11\}$$

presenta moda pari a $Mo = 2$ perché tale valore è quello più frequente (si presenta 4 volte). Eppure, tale moda appare difficilmente «rappresentativa» dell'intera distribuzione di frequenza. Infatti, le modalità 7 e 9 compaiono 3 volte; inoltre, la maggior parte della distribuzione (i $\frac{13}{20}$ delle unità, cioè il 65%) assume valore superiore a 6, e pochi sarebbero propensi a sintetizzare il fenomeno assumendo come sintesi rappresentativa il valore 2.

Quest'ultimo caso mostra che, in effetti, la moda assolve la sua funzione di rappresentare la variabile quando la distribuzione di frequenza è sicuramente unimodale e la moda non è troppo distante dai valori centrali della distribuzione. □

Esempio 4.9

I dati della tabella seguente (tratti dal Censimento ISTAT del 1991) evidenziano la distribuzione di frequenza per la *variabile qualitativa nominale* «Stato civile» dei cittadini residenti in Italia nelle modalità: nubile/celibe, coniugata/o, separata/o legalmente, divorziata/o, vedova/o.

Stato civile	Frequenze assolute	relative	percentuali
Nubile/Celibe	23.515.646	0.41417	41.417
Coniugata/o	28.184.704	0.49640	49.640
Separata/o	478.537	0.00843	0.843
Divorziata/o	375.369	0.00661	0.661
Vedova/o	4.223.775	0.07439	7.439
Totali	56.778.031	1.00000	100.000

La modalità più frequente è coniugata/o che, quindi, è la moda della distribuzione. Questo implica che – dal punto di vista dello stato civile – la situazione della maggior parte della popolazione residente è quella di essere sposati. □

o Un errore registrato con sorprendente frequenza tra gli studenti è quello di calcolare la moda come il massimo della variabile e non come quella modalità cui corrisponde la massima frequenza. Così, accade spesso di ascoltare che, con le osservazioni {8, 1, 1, 2, 4}, la modalità 8 è la moda! invece, 8 è la modalità massima della variabile, non la modalità più frequente che è «1» perché si verifica 2 volte (ossia più di qualsiasi altra modalità).

Un altro errore è quello di confondere la moda con la frequenza assoluta massima. Per evitarlo, basti pensare che nel linguaggio corrente si intende per «moda» un comportamento, un vestito, un oggetto, una canzone, e non la frequenza delle persone che adotta quel comportamento, indossa quel vestito, compra quell'oggetto, preferisce quella canzone.

Come tutti gli indici di posizione, la moda deve essere espressione del fenomeno oggetto di studio: la moda delle altezze in cm sarà un'altezza espressa in cm; la moda del reddito misurato in migliaia di Euro/mese sarà un reddito espresso in migliaia di Euro/mese, e così via. Quindi, essa deve essere un valore scelto tra quelli manifestati dalla variabile nella popolazione: infatti, *la moda è una modalità, non una frequenza.*

Per variabili continue, appare difficile giustificare la individuazione di un solo valore della variabile come moda della distribuzione. Tuttavia, volendo procedere

comunque al calcolo della moda per variabili raggruppate in classi, occorre individuare la classe di modalità cui corrisponde la massima frequenza, per cui si parla di *classe modale*.

A tal fine, occorre prestare attenzione al fatto che se le classi non sono equi-ampie non è corretto confrontare frequenze assolute (e relative) che fanno riferimento ad intervalli di modalità della variabile di ampiezza differenti. In tali casi, occorre dividere la frequenza assoluta di ciascuna classe per l'ampiezza dell'intervallo cui si riferisce pervenendo alla distribuzione di quella che si chiama «*densità di frequenza*»: pertanto, *la classe modale è quella con densità di frequenza più elevata.*

Esempio 4.10

Per il data set COMPACT, esaminiamo la variabile «Durata in secondi» dei 1192 brani musicali. Per tale variabile otteniamo la distribuzione di frequenza della tabella seguente, nella quale le classi non sono equi-ampie ma sono state costruite in modo che avessero una interpretazione immediata (per esempio fino a 1 minuto, da 3 a 6 minuti, etc.)

Classi di modalità	Frequenze assolute	Frequenze relative	Densità di frequenza
30 ⊢ – ⊣ 60	5	0.00419	0.16667
60– ⊣ 150	133	0.11158	1.47778
150– ⊣ 180	149	0.12500	4.96667
180– ⊣ 300	571	0.47903	4.75833
300– ⊣ 480	284	0.23826	1.57778
480– ⊣ 600	28	0.02349	0.23333
600– ⊣ 900	21	0.01762	0.07000
900– ⊣ 1022	1	0.00084	0.00820

Ebbene, non potremo affermare che la classe modale è quella con maggiore frequenza assoluta (o relativa), cioè (180,300] perché le classi non sono equi-ampie per cui nel confronto occorre tenere conto anche dell'ampiezza delle classi. L'ultima colonna, ottenuta come rapporto tra la frequenza assoluta delle classi e la loro ampiezza esprime una sorta di densità media di ciascuna classe, per cui è in effetti (150,180] la classe modale per la variabile Durata.

Problemi simili sono comuni con variabili continue, come la distribuzione dei Redditi IRPEF dichiarati in Italia, i Consumi familiari, la Superficie delle aziende agrarie, etc. □

■ Un indice di posizione sicuramente più rappresentativo è la *mediana* Me, che è definita come quel valore di una variabile X che bipartisce la distribuzione ordinata delle modalità $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Quindi, la mediana è determinata in modo che metà delle osservazioni siano inferiori alla mediana e metà siano ad essa superiori.

In breve, *la mediana è la modalità della unità statistica che occupa il posto centrale nella distribuzione ordinata delle osservazioni.*

○ Si osservi che la mediana è calcolabile per ogni variabile (qualitativa o quantitativa) per la quale sia possibile stabilire un ordinamento tra le modalità; in particolare, essa si può calcolare per variabili ordinali.

○ Ricordando la definizione della funzione di ripartizione, si vede subito che la mediana è quel valore per il quale la funzione di ripartizione vale $\frac{1}{2}$, cioè: $F(Me) = \frac{1}{2}$.

○ Operativamente, conviene distinguere tra variabili discrete con dati singolarmente raccolti o già organizzati in una distribuzione di frequenza e variabili continue con classi di modalità e rispettive frequenze.

• *Per variabili X discrete*, registrate singolarmente e ordinate dal valore minimo $x_{(1)}$ al valore massimo $x_{(n)}$, la mediana Me è quel valore così definito:

$$Me = \begin{cases} x_{(\frac{n+1}{2})}, & \text{se } n \text{ è dispari;} \\ \dfrac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}, & \text{se } n \text{ è pari.} \end{cases}$$

La formula, quindi, definisce la mediana come quell'unico valore centrale per dati ordinati se questi sono in numero dispari, ovvero come la semi-somma dei due valori centrali se questi sono in numero pari.

○ In effetti, a voler essere rigorosi, quando i dati sono in numero pari la mediana è qualsiasi valore compreso tra i due valori centrali. Tuttavia, agli effetti pratici, la precedente convenzione aiuta nella determinazione di un unico valore numerico per la mediana.

Esempio 4.11

○ Una variabile X in una popolazione di 21 unità statistiche presenta i seguenti valori, già ordinati dal minimo al massimo:

$$\{1, 2, 2, 2, 2, 3, 4, 6, 6, 7, 7, 7, 8, 8, 9, 9, 9, 10, 11, 11, 15\}.$$

Essendo le modalità in numero dispari, la mediana sarà la modalità collocata al posto $\frac{n+1}{2} = \frac{22}{2} = 11$, cioè all'11-esimo posto della graduatoria. Ora, nella popolazione ordinata, all'11-esimo posto vi è la modalità 7 cioè $x_{(11)} = 7$, ed allora: $Me = 7$.

○ Invece, per la popolazione seguente:

$$\{-10, -7, -7, -3, -3, -2, -2, 0, 0, 1, 3, 7, 8, 9, 9, 10, 12, 16, 17, 23\},$$

che consiste di un numero pari di 20 unità, già ordinate dal minimo al massimo, la mediana può essere definita come la media aritmetica (cioè la semi-somma) dei due valori di posto centrale. Ora, i due valori centrali saranno collocati, rispettivamente, al posto $\frac{n}{2} = \frac{20}{2} = 10$ e al posto $(\frac{n}{2} + 1) = (\frac{20}{2} + 1) = 11$, cioè al 10-imo (decimo) e all'11-esimo (undicesimo) posto. Ora, nella popolazione ordinata al decimo posto e all'undicesimo posto vi sono i valori, rispettivamente: $x_{(10)} = 1$ e $x_{(11)} = 3$, per cui la mediana è: $Me = \frac{1+3}{2} = 2$.

□

• *Per variabili X discrete*, note mediante una seriazione (perché organizzate in una distribuzione di frequenza), occorre individuare la modalità dell'unità statistica corrispondente al posto centrale, il che è generalmente semplice se i dati sono stati ordinati per calcolare le frequenze relative cumulate.

Infatti, dalla conoscenza della funzione di ripartizione $F(x)$, si individua la modalità x_{i-1} cui corrisponde $F(x_{i-1}) < 0.5$ e tale che per la modalità successiva x_i si abbia $F(x_i) \geq 0.5$. Evidentemente, la mediana Me coincide con la modalità x_i perché tra le n_i unità che possiedono modalità x_i vi sarà certamente quella (se n è dispari) o quelle (se n è pari) di posto centrale.

Esempio 4.12

o Riprendiamo i dati esposti nell'esempio 3.6 mediante la distribuzione di frequenza ivi riportata. Si vede subito che in corrispondenza di $X = 10$ la funzione di ripartizione vale $F(10) = 0.40$ mentre in corrispondenza di $X = 12$ la funzione di ripartizione vale $F(12) = 0.65$. Ciò implica che il 40% delle unità statistiche presenterà valori delle modalità inferiori a 10 e il 65% delle unità statistiche presenterà valori delle modalità inferiori a 12. Ne consegue che vi saranno due unità statistiche (perché in questo esempio la numerosità è pari, essendo $n = 20$) al centro della distribuzione che, entrambe, hanno modalità pari a 12. Quindi, la mediana è $Me = 12$.

	Frequenze *Maschi*			Frequenze *Femmine*		
Modalità	assolute	relative	f.ripartizione	assolute	relative	f.ripartizione
34	0	0.00000	0.00000	2	0.00163	0.00163
35	0	0.00000	0.00000	35	0.02846	0.03008
36	0	0.00000	0.00000	148	0.12033	0.15041
37	0	0.00000	0.00000	327	0.26585	0.41626
38	4	0.00515	0.00515	359	0.29187	0.70813
39	12	0.01544	0.02059	227	0.18455	0.89268
40	53	0.06821	0.08880	115	0.09350	0.98618
41	107	0.13771	0.22651	13	0.01057	0.99675
42	212	0.27284	0.49936	3	0.00244	0.99919
43	178	0.22909	0.72844	1	0.00081	1.00000
44	123	0.15830	0.88674	0	0.00000	1.00000
45	68	0.08752	0.97426	0	0.00000	1.00000
46	18	0.02317	0.99743	0	0.00000	1.00000
47	1	0.00129	0.99871	0	0.00000	1.00000
48	0	0.00000	0.99871	0	0.00000	1.00000
49	1	0.00129	1.00000	0	0.00000	1.00000
Totali	777	-----	-----	1230	-----	-----

o Dalla funzione di ripartizione delle famiglie residenti in Lombardia al Censimento 1981 (di cui all'esempio 3.7) risulta $F(2) = 0.4332$; $F(3) = 0.6758$. Pertanto, se il 43.32% di tali famiglie ha 2 componenti, mentre il 67.58% di tali famiglie ha 3 componenti, allora supponendo idealmente allineate le famiglie (prima quelle con 1 componente, poi quelle con 2 componenti, poi quelle con 3 componenti, etc.), vi sarà certamente una famiglia con 3 componenti che sarà al centro della distribuzione. Quindi, metà delle famiglie avrà un numero di componenti minore o uguale a 3 e metà delle famiglie avrà un numero di componenti maggiore o uguale a 3. Allora, la mediana è $Me = 3$.

Tale famiglia «mediana» bipartisce la distribuzione delle famiglie e, quindi, la corrispondente modalità $(X = 3)$ della variabile X sarà la mediana della distribuzione.

o Dai dati dell'esempio 4.2, che riprendiamo nella tabella sopra riportata assieme alle frequenze relative e alla funzione di ripartizione empirica, deduciamo che la mediana per la variabile «Misura delle scarpe» è 43 per i Maschi e 38 per le Femmine.

Esaminando la funzione di ripartizione empirica per i maschi, emerge che il valore 0.5 si riscontra in corrispondenza della modalità 43 per cui la mediana è 43. Infatti, ordinando le modalità dichiarate dai 777 studenti, al centro vi sarà la modalità di posto 389-esimo e dalla tabella si vede che tale studente ha dichiarato una misura pari a 43 (si osservi che, in questo particolare esempio, lo studente al 388-esimo posto ha dichiarato una misura pari a 42). Quindi, $Me = 43$.

Similmente, esaminando la funzione di ripartizione empirica per le femmine, emerge che il valore 0.5 si riscontra in corrispondenza della modalità 38 per cui la mediana è 38. Infatti, ordinando le modalità dichiarate dalle 1230 studentesse, al centro vi saranno le modalità di posto 615-esimo e 616-esimo e dalla tabella si vede che tali studentesse hanno entrambe dichiarato una misura pari a 38. Quindi, $Me = 38$. □

• *Per variabili X continue*, il raggruppamento in classi delle modalità consente al più di determinare solo una *classe mediana* nella quale ricade l'unità statistica che bipartisce la distribuzione ordinata delle modalità.

o Un singolo indice sintetico può essere ottenuto approssimando la funzione di ripartizione attorno alla mediana. Il modo più semplice è quello di ipotizzare una approssimazione lineare (cioè mediante una retta) che conduce alla determinazione della mediana mediante la formula seguente:

$$Me \simeq x_{i-1} + (x_i - x_{i-1}) \frac{0.5 - F(x_{i-1})}{F(x_i) - F(x_{i-1})}$$

nella quale abbiamo supposto che: $F(x_{i-1}) < 0.5$; $F(x_i) \geq 0.5$.

Esempio 4.13

Per il data set COMPACT, con le informazioni presentate nell'esempio 3.8, sappiamo che la funzione di ripartizione empirica della variabile X = «Durata in secondi» dei brani musicali assume i valori:

$$F(210.364) = 0.35403 < 0.5; \quad F(300.545) = 0.71980 > 0.5.$$

Quindi, la classe mediana sarà (210.364, 300.545] per cui la mediana sarà sicuramente collocata in tale intervallo, cioè:

$$210.364 \leq Me \leq 300.545.$$

Se interessa determinare un solo valore per la mediana, si può applicare la formula approssimata dalla quale si avrà:

$$Me \simeq 210.364 + (300.545 - 210.364) \frac{0.5 - 0.35403}{0.71980 - 0.35403} = 246.35307.$$

Ora, per il data set COMPACT si conoscono le singole osservazioni per cui, ordinando le modalità in modo non-decrescente, è possibile pervenire alla determinazione esatta della mediana. Infatti, poiché la popolazione è di $n = 1192$ osservazioni, la mediana è qualunque valore reale compreso tra le modalità di posto centrale, cioè tra quelle registrate sul 596-esimo $\left(= \frac{1192}{2}\right)$ e il 597-esimo $= \left(\frac{1192}{2} + 1\right)$ brano. Ora, sia il 596-esimo che il 597-esimo brano hanno durata uguale perché: $x_{596} = x_{597} = 244$ secondi, per cui: $Me = 244$ secondi.

In sostanza, la mediana esprime il fatto che il 50% dei brani dura al più 244 secondi (4 minuti e 4 secondi) e il 50% dura non meno di 244 secondi.

Si osservi come, per questo data set, l'approssimazione lineare sia risultata particolarmente soddisfacente.

□

La mediana gode di alcune proprietà tra le quali va ricordato che, a differenza della media, la mediana è sempre un valore realmente osservato nella popolazione in esame.

In secondo luogo, *la mediana è l'unico valore che minimizza la somma degli scarti presi in valore assoluto.* Più precisamente, la quantità: $\sum_{i=1}^{n} |x_i - c|$ è minima se e solo se $c \equiv Me$.

Tale proprietà ha conseguenze nella scelta dell'allocazione di un deposito se si desiderano minimizzare i costi di trasporto tra i vari punti di vendita al dettaglio e il suddetto deposito. Supponendo, per semplicità, che i costi siano proporzionali alla distanza e che i vari punti vendita siano posizionati ai km indicati dalla sequenza $\{x_1, x_2, \ldots, x_n\}$, il deposito andrà sistemato in un punto c in modo da minimizzare la somma di tutti i costi necessari per il rifornimento, cioè occorrerà minimizzare $\sum |x_i - c|$; la citata proprietà suggerisce di collocare il deposito nella mediana dei punti in cui sono posizionati i punti vendita.

Tuttavia, la proprietà sicuramente più importante è la capacità della mediana di essere rappresentativa della posizione della distribuzione anche in presenza di valori estremi notevolmente diversi da tutti gli altri: tale requisito è detto *resistenza* e costituisce un indubbio vantaggio della mediana rispetto alla media aritmetica.

- La resistenza a valori estremi deriva dal fatto che il calcolo della mediana tiene conto solo dell'ordinamento delle osservazioni, limitandosi a considerare la modalità dell'elemento collocato al centro della graduatoria ordinata.
- Naturalmente, proprio perché determinata dalla modalità centrale, la mediana è estremamente sensibile alle modifiche indotte nel corpo centrale della distribuzione. Quindi, va utilizzata con estrema cautela quando la differenza tra due popolazioni è rilevante proprio nel centro della distribuzione ordinata delle modalità.

Esempio 4.14

La tabella seguente mostra tre popolazioni (espresse già secondo modalità ordinate, per convenienza di calcolo) che differiscono tra loro solo per la presenza di un unico valore. Per entrambe sono calcolate ed evidenziate sia la media che la mediana.

Popolazione	Modalità osservate	Media	Mediana
A	{1, 3, 12, 24, 25, 26, 27, 35, 36, 38, 40}	24.273	26
B	{1, 3, 12, 24, 25, 26, 27, 35, 36, 38, 400}	57.000	26
C	{−100, 3, 12, 24, 25, 26, 27, 35, 36, 38, 40}	11.545	26

Come si vede, mentre la mediana non viene alterata dalla presenza di valori estremi (nel caso B eccezionalmente alto, nel caso C eccezionalmente basso), la media viene drasticamente spostata nella direzione dell'elemento eccezionale.

□

Esempio 4.15

Supponiamo che si intenda riassumere la preparazione universitaria di 4 studenti X, Y, Z, W che alla fine del loro percorso di 19 esami abbiano conseguiti i voti presentati nelle prime due righe della tabella seguente:

Studente	Modalità osservate	Media	Mediana
X	{18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 30, 30, 30, 30, 30, 30, 30, 30, 30}	23.684	18
Y	{18, 18, 18, 18, 18, 18, 18, 18, 18, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30}	24.316	30
Z	{18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18}	18.000	18
W	{30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30}	30.000	30

I primi due si differenziano per un solo voto, e la media evidenzia tale situazione in modo corretto esplicitando una posizione di X leggermente inferiore di circa mezzo voto rispetto a quella di Y.

Per contro, la mediana – essendo derivata dalla sola modalità dell'esame di posto centrale, il decimo in tale esempio – ha un effetto devastante nel determinare la posizione dei due studenti conferendo ad X la stessa posizione dello studente Z, che ha la preparazione minima ammissibile in teoria. Similmente, la mediana assimila la preparazione dello studente Y a quella dello studente W, la cui preparazione è la massima ammissibile in teoria.

Sarebbe, quindi, profondamente ingiusto utilizzare la mediana come sintesi della preparazione universitaria adducendo il principio che, essendo la mediana resistente, essa limita gli effetti di valori atipici sulla sintesi. Un principio del genere indurrebbe gli studenti a studiare solo la metà più 1 degli esami da sostenere, essendo infatti la mediana determinata solo dal valore centrale della distribuzione.

□

L'ultimo esempio conferma il principio generale secondo il quale *la sintesi di una variabile va determinata sulla base degli obiettivi da perseguire*, per cui:

i) per valutare la preparazione complessiva occorre tenere conto di tutti i valori e, quindi (secondo il criterio della sostituzione del Chisini), la *media aritmetica* è l'indicatore di posizione da utilizzare;

ii) se in un'asta pubblica si vuole eliminare l'effetto artificioso di concorrenti sleali, che suggeriscono prezzi anomali per modificare la distribuzione delle offerte, occorre puntare su indicatori resistenti, e quindi la *mediana* è l'indicatore di posizione da utilizzare.

4.7 Quantili di una distribuzione di frequenza

Come si è visto, la mediana lascia a destra e a sinistra del suo valore il 50% delle modalità, suddividendo esattamente a metà la numerosità delle unità comprese tra il minimo e la mediana e, ovviamente, tra la mediana e il massimo.

Se si opera in modo analogo, ma suddividendo in quattro gruppi di pari numerosità la popolazione di n unità (sempre ordinate in senso non-decrescente), si individueranno 3 valori, definiti ***quartili***, così determinati:

- Q_1, detto *primo quartile*, è il valore della modalità della variabile X tale che al di sotto di Q_1 vi saranno $\frac{1}{4}n$ unità statistiche (cioè il 25% delle unità statistiche presenta valori inferiori a Q_1);
- Q_2, detto *secondo quartile*, è il valore della modalità della variabile X tale che al di sotto di Q_2 vi saranno $\frac{2}{4}n$ unità statistiche (cioè il 50% delle unità statistiche presenta valori inferiori a Q_2);
- Q_3, detto *terzo quartile*, è il valore della modalità della variabile X tale che al di sotto di Q_3 vi saranno $\frac{3}{4}n$ unità statistiche (cioè il 75% delle unità statistiche presenta valori inferiori a Q_3).

Si comprende che *il secondo quartile coincide con la mediana*: $Q_2 \equiv Me$.

È possibile estendere tali suddivisioni di pari numerosità, tra il minimo ed il massimo delle modalità osservate nella popolazione: si parla così di decili, percentili e, in generale, di ***quantili*** di una distribuzione di frequenza.

o Per esempio, i ***decili*** sono i valori D_1, D_2, ..., D_9 tali che al di sotto di essi vi siano, rispettivamente, il 10%, il 20%,..., il 90% delle unità statistiche.

Esempio 4.16

Supponiamo che due studenti, indicati con A e B, abbiano riportati in 20 esami del proprio curriculum di studi i voti seguenti:

$A \rightarrow$	18	19	19	19	19	20	20	20	21	21	22	22	22	22	23	24	24	24	25	25
$B \rightarrow$	18	22	23	23	23	25	25	25	27	27	27	27	27	28	28	28	28	30	30	30

Essendo $n = 20$, ed avendo già ordinato le modalità della variabile Voto, i tre quartili Q_1, Q_2, Q_3 si troveranno ai posti: $\frac{1}{4}n = 5$; $\frac{2}{4}n = 10$; $\frac{3}{4}n = 15$, cioè ai posti 5°, 10°, 15° rispettivamente.

Pertanto, per i due studenti si ha:

Studenti	Q_1	Q_2	Q_3
$A \rightarrow$	19	21	23
$B \rightarrow$	23	27	28

Emerge con chiarezza che la distribuzione dei voti dello studente B è più spostata verso valori alti rispetto a quella dello studente A. □

Esempio 4.17

La variabile X = «Durata in secondi» per il data set COMPACT consiste di 1192 brani che variano da un minimo di $x_{(1)} = 30$ ad un massimo di $x_{(n)} = 1022$ secondi.

Pertanto, dopo aver ordinato le durate in modo non-decrescente, essendo $\frac{n}{4} = 298$, il primo quartile sarà la semisomma delle durate dei brani collocati al 298-esimo e 299-esimo posto nella graduatoria; poiché risulta: $x_{(298)} = x_{(299)} = 182$, allora il primo quartile sarà $Q_1 = 182$ secondi.

Similmente, poiché $2\frac{n}{4} = 596$, il secondo quartile (che coincide con la mediana) sarà la semisomma delle durate dei brani collocati al 596-esimo e al 597-esimo posto; poiché risulta: $x_{(596)} = x_{(597)} = 244$, allora il secondo quartile sarà $Q_2 = Me = 244$ secondi.

Infine, poiché $3\frac{n}{4} = 894$, il terzo quartile sarà la semisomma delle durate dei brani collocati all'894-esimo e all'895-esimo posto; poiché risulta: $x_{(894)} = x_{(895)} = 308$, allora il terzo quartile sarà $Q_3 = 308$ secondi.

Pertanto, riassumendo, la variabile X è caratterizzata dai seguenti valori:

$$x_{(1)} = 30;\ Q_1 = 182;\ Q_2 = 244;\ Q_3 = 308;\ x_{(n)} = 1022;$$

Emerge chiaramente che la distribuzione possiede valori estremi rispetto al suo centro. □

Una delle ragioni principali per cui mediana, quantili e moda sono delle sintesi utili della posizione di una distribuzione di frequenza è che esistono sempre all'interno della distribuzione in esame. Infatti, essi sono i valori assunti dalle modalità della variabile in corrispondenza di una unità statistica che occupa una prefissata posizione o che possiede una determinata caratteristica. Invece, la media è il risultato di calcoli aritmetici che possono dar luogo a valori che soddisfano proprietà formali ma che non necessariamente si ritrovano nei casi osservati.

o Per esempio, se i $\frac{2}{3}$ di un collettivo possiede un reddito mensile di 4000 Euro e $\frac{1}{3}$ possiede un reddito mensile di 1000 Euro, la media aritmetica è di 3000 Euro: ma nessuna unità statistica possiede tale reddito! Invece, la mediana e la moda sono entrambe pari a 4000 Euro e rappresentano, rispettivamente, il reddito mensile dell'unità di posto centrale (nel caso della mediana) e il reddito mensile della maggior parte della popolazione (nel caso della moda).

o Similmente, il voto mediano è un voto che lo studente ha effettivamente ricevuto all'esame collocato al centro della graduatoria non-decrescente di tutti i voti degli esami;

la moda delle età in anni compiuti è l'età più diffusa tra i componenti di un certo gruppo, e così via.

Prima di concludere questo paragrafo, riassumiamo i possibili indici di posizione che possono essere calcolati per variabili statistiche in funzione della loro natura.

- La *moda* si può calcolare per tutte le tipologie di variabili.
- La *mediana*, i *quartili* e, in generale, i *quantili* si possono calcolare per tutte le variabili le cui modalità possono essere ordinate (quindi, per *variabili qualitative ordinali*, oltre che per tutte le variabili quantitative).
- La *media* si può calcolare solo per variabili quantitative.

Va precisato che, nel caso di variabili qualitative, potrebbe sorgere il problema di calcolare la semisomma dei valori di posto centrale quando i dati sono in numero pari, il che è concettualmente impossibile trattandosi di attributi e non di numeri. In queste situazioni occorre procedere sulla base di altri criteri per scegliere fra due attributi contigui.

4.8 Altri indici di posizione

La letteratura statistica si è arricchita di numerosi altri indici di posizione, sia specificando in modo diverso la funzione $f(\ldots)$ presente nel criterio di sostituzione di Chisini che cercando misure che siano più resistenti della media senza perdere alcune delle sue proprietà fondamentali.

Sul primo versante, accenniamo alla media geometrica e alla media armonica, il cui uso (soprattutto per la seconda) è piuttosto circoscritto; sul secondo versante presenteremo alcune medie troncate che, in talune situazioni, potrebbero risultare utili.

- La ***media geometrica*** deriva dal criterio di Chisini definendo la funzione $f(x_1, x_2, \ldots, x_n)$ come il prodotto delle modalità $(x_1, x_2, \ldots, x_n)$. Dopo agevoli passaggi, si deduce che la media geometrica assume significato solo se tutte le modalità sono strettamente positive e, in tal caso, è definita da:

$$\mathbb{G} = \sqrt[n]{x_1\, x_2 \ldots x_n} = \sqrt[n]{x_1^{n_1}\, x_2^{n_2} \ldots x_k^{n_k}}, \quad \forall x_i > 0,$$

dove la prima espressione va applicata se i dati sono disponibili in modalità distinte (serie) e la seconda se sono disponibili in seriazione (cioè mediante una distribuzione di frequenza).

- Dal punto di vista computazionale, la media geometrica può essere utilmente calcolata mediante i logaritmi, cioè utilizzando una delle due seguenti espressioni:

$$\mathbb{G} = \exp\left\{\frac{1}{n}\sum_{i=1}^{n}\log(x_i)\right\} = \exp\left\{\frac{1}{n}\sum_{i=1}^{k}\log(x_i)\, n_i\right\}$$

in funzione della disponibilità dei dati in serie (per la prima) o in seriazione (per la seconda).

- Come si vede, la media geometrica è l'esponenziale della media aritmetica del logaritmo (neperiano) delle modalità.

- La ***media armonica*** deriva dal criterio di Chisini definendo la funzione $f(x_1, x_2, \ldots, x_n)$ come la somma degli inversi delle modalità $(x_1, x_2, \ldots, x_n)$. Dopo qualche passaggio, si deduce che la media armonica è definita da:

$$\mathbb{A} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}} = \frac{n}{\sum_{i=1}^{k}\frac{n_i}{x_i}},$$

dove la prima espressione va applicata se i dati sono disponibili in modalità distinte (serie) e la seconda se sono disponibili in seriazione (cioè mediante una distribuzione di frequenza).

Come si vede, la media armonica è l'inverso della media aritmetica degli inversi delle modalità.

- Per dati strettamente positivi, si dimostra che:

$$min(x_i) \leq \mathbb{A} \leq \mathbb{G} \leq \mu \leq max(x_i).$$

Ovviamente, le tre medie coincidono se e solo se tutte le modalità coincidono, cioè se la variabile X assume un valore costante per tutte le unità della popolazione: in tal caso, la distribuzione si chiama *degenere*.

- Le definizioni precedenti sono state date per una variabile discreta. Quando, invece, la variabile è nota mediante modalità raggruppate in classi, con le corrispondenti frequenze, si conviene di utilizzare il valore centrale della classe, discretizzando di fatto la variabile continua. Tale convenzione è del tutto arbitraria ed è giustificata solo dalla sua semplicità.

Esempio 4.18

Per una variabile quantitativa X che presenta le seguenti modalità:

$$\{\ 1,\ 3,\ 2,\ 5,\ 4,\ 6,\ 8,\ 10,\ 20,\ 10\ \}$$

i calcoli per ottenere le medie aritmetica, geometrica e armonica sono i seguenti:

$$\mu = \frac{1+3+2+\ldots+10}{10} = \frac{69}{10} = 6.90000$$

$$\mathbb{G} = \exp\left\{\frac{\log(1)+\log(3)+\log(2)+\ldots+\log(10)}{10}\right\} = \exp\left\{\frac{16.25959}{10}\right\} = 5.08329$$

$$\mathbb{A} = \frac{10}{\frac{1}{1}+\frac{1}{3}+\frac{1}{2}+\ldots+\frac{1}{10}} = \frac{10}{2.825} = 3.53982$$

Come si vede, è confermato che tutte le medie sono comprese tra il minimo e il massimo delle osservazioni, e che la media armonica è non superiore alla media geometrica la quale, a sua volta, è non superiore alla media aritmetica.

□

Ulteriori proposte di medie sono state avanzate soprattutto per limitare i danni provocati dai valori estremi (perché realmente eccezionali oppure perché verosimilmente errati) e tra queste ci limitiamo a segnalare quelle più semplici.

- Una ***media troncata*** al 2.5% è una media aritmetica di tutte le modalità ordinate dopo aver eliminato dalla popolazione il 2.5% dei valori più bassi e il 2.5% dei valori più alti. Quindi, tale media è ottenuta calcolando la media aritmetica del 95% della popolazione di posto centrale nella serie ordinata delle osservazioni.

  - In funzione del tasso di dati anomali presenti si può ridurre la detta percentuale (per inserire più dati) oppure ampliarla (per eliminare più dati estremi). È agevole convincersi che una media troncata al 50% è la mediana.

- Una soluzione intermedia è la cosiddetta ***media secondo Winsor*** la quale anziché eliminare una porzione dei dati estremi li sostituisce con il dato più vicino, come è mostrato nell'esempio seguente.

Esempio 4.19

Confrontiamo la media aritmetica e la media troncata al 5% (per la quale, quindi, si eliminano complessivamente il 10% delle osservazioni estreme) nella seguente popolazione di $n = 40$ osservazioni:

$$0,1,8,9,14,14,14,25,25,27,\quad 38,38,39,40,40,40,41,44,44,45$$
$$47,48,49,50,50,50,50,52,53,55,\quad 61,63,64,67,70,70,72,74,120,200$$

In tal caso, per ottenere la media troncata μ_T al 5%, si tratta di ordinare le osservazioni e di calcolare la media aritmetica dopo aver eliminato il 5% (= 2 osservazioni) delle modalità più piccole e il 5% (= 2 osservazioni) delle modalità più grandi.

Invece, per ottenere la media secondo Winsor μ_W occorre sostituire le prime 2 osservazioni con il valore 8 e le ultime 2 con il valore 74.

Per le tre medie (aritmetica, troncata, secondo Winsor), si ottengono i seguenti valori:

$$\mu = \frac{0+1+8+9+14+\ldots+72+74+120+200}{40} = \frac{1911}{40} = 47.77500$$

$$\mu_T = \frac{8+9+14+\ldots+72+74}{36} = \frac{1590}{36} = 44.16667$$

$$\mu_W = \frac{8+8+8+9+14+\ldots+72+74+74+74}{40} = \frac{1754}{40} = 43.85000$$

Come è evidente, la media aritmetica viene alterata (in questo esempio) soprattutto da valori estremi elevati, mentre le altre due medie ri-collocano la misura di posizione più al centro della distribuzione.

Si noti che, per questi dati, la mediana, ottenuta come la semi-somma delle modalità collocate al 20-esimo e al 21-esimo posto, vale: $Me = 46$.

□

4.9 Considerazioni finali

In questo Capitolo sono stati introdotti alcuni indici di posizione che avevano tutti l'obiettivo di riassumere in modo rappresentativo la variabile di interesse.

Tale sintesi è stata raggiunta, in primo luogo, mediante il criterio di sostituzione pervenendo alle medie aritmetica, geometrica e armonica.

Per contro, osservando la particolare collocazione di specifici elementi della distribuzione si è pervenuti ad altre misure, come i quantili (in particolare, la mediana e i quartili) e la moda.

Infine, alcuni indici di posizione sono stati derivati per tener conto degli aspetti positivi (completezza e resistenza ai valori estremi) presenti nei due precedenti criteri (ciò vale per le medie troncate e secondo Winsor).

Ora, anche se va riaffermato il principio generale secondo cui saranno gli obiettivi a far propendere la scelta verso un indice tra quelli proposti, è bene indicare quali obiettivi orientano verso un indice statistico di posizione. Lo faremo, soffermandoci su moda, mediana e media che costituiscono di gran lunga i più diffusi.

- La ***moda*** è utile quando occorre «minimizzare gli scontenti», e quindi in tutte quelle situazioni ove il consenso ed il numero delle singole unità ha significato per la decisione. In breve, *la moda è un indice utile per individuare la modalità più rappresentativa.*

- La ***mediana*** «minimizza i costi complessivi» ed è soprattutto resistente a valori estremi. In breve, *la mediana è un indice per decisioni che implicano costi elevati nei casi estremi.*

- La ***media*** aritmetica è il baricentro dei dati e quindi propone un valore che equi-ripartisce il fenomeno tra le unità statistiche, pervenendo così a decisioni nelle quali contano – a parità numerica – gli estremi molto più dei valori centrali. In breve, *la media aritmetica è un indice di equilibrio generale.*

In conclusione, per determinare un valore sintetico della posizione di una distribuzione di frequenza di una variabile quantitativa X, si può suggerire di calcolare almeno la media e la mediana e dal loro confronto – come vedremo più avanti – dedurre utili informazioni sia sulla posizione che sulla forma della distribuzione.

Ambiente R

Indici statistici di posizione

R4.1 Media aritmetica

La Statistica utilizza innumerevoli indici di posizione che, di volta in volta, sono efficaci per casi particolari e qui ci limitiamo ad illustrare il calcolo in R di quelli definiti ed esemplificati nel testo.

La funzione `mean(x)` calcola la ***media aritmetica*** delle modalità quantitative presenti in x. Se x è un vettore, allora `mean(x)` restituisce la media di tale variabile; se x è una matrice, allora `mean(x)` restituisce, ordinatamente, la media delle variabili presenti nelle colonne di `dati`.

Il data set `camellia` contiene (alle colonne 2, 3, 4, 5, 6) le variabili `lunghezza`, `larghezza`, `pesovivo`, `pesosecco` (che sono quantitative), e la variabile `codice` (che è dicotomica).

```
R==> dati=read.table("camellia.txt", header=TRUE);
R==> lunghezza=dati[,2]; variabili=dati[,2:6];

R==> mean(lunghezza);
[1] 97.565

R==> mean(variabili);
 lunghezza  larghezza   pesovivo  pesosecco     codice
  97.56456   45.71772 9985.21321 3961.53453    0.26126
```

Quindi, la lunghezza media delle foglie è di 97.565 mm mentre la larghezza media è di 45.718 mm. Si noti che la media aritmetica della variabile `codice` (dicotomica, che vale 1 per la specie Lanceolata), è in effetti la somma delle modalità pari ad 1, divisa per la numerosità dei dati; quindi, esprime la frazione di foglie che appartengono alla specie Lanceolata (nello specifico, il 26%).

Quando si dispone di una matrice o di una data frame è conveniente operare la media delle variabili con la funzione `colMeans()`, come nell'esempio:

```
R==> dati=read.table("camellia.txt", header=TRUE);
R==> colMeans(dati[,2:6]);
 lunghezza  larghezza   pesovivo  pesosecco     codice
  97.56456   45.71772 9985.21321 3961.53453    0.26126
```

In presenza di dati mancanti, la media non si applica a meno di non specificare l'opzione `na.rm=TRUE` (che significa "rimuovi NA"). Per esempio:

```
R==> esempio=c(1,3,2,4,NA,4,5,NA,6,NA);
R==> mean(esempio)
[1] NA

R==> mean(esempio, na.rm=TRUE);
[1] 3.5714
R==> mean(c(1,3,2,4,4,5,6));
[1] 3.5714
```

Se i dati sono disponibili in seriazione (cioè sotto forma di modalità e frequenze mediante due corrispondenti vettori `xi` e `ni`, per esempio), il calcolo della media aritmetica avviene nel modo seguente:

```
R==> xi=c(1,2,3,5,7);  ni=c(10,12,8,2,26);
R==> media=sum(xi*ni)/sum(ni);
R==> media;
[1] 4.3103
```

Come si vede, la formula $\frac{\sum_i x_i n_i}{\sum_i n_i}$ è stata tradotta letteralmente in R grazie al suo formalismo vettoriale. Ovviamente, si potrebbe anche replicare la modalità 1 per 10 volte, la modalità 2 per 12 volte, ..., la modalità 7 per 26 volte e poi applicare la funzione `mean()` alla serie così ottenuta: questo è agevole (ma non efficiente) in R mediante le istruzioni:

```
### seriazione, composta di modalità e frequenze
R==> xi=c(1,2,3,5,7);  ni=c(10,12,8,2,26);
### ricostruzione della serie tramite rep()
R==> serie=rep(xi,ni);   # ripete x_i per n_i volte
R==> mean(serie);
[1] 4.3103
```

In modo analogo, si calcola la ***media aritmetica ponderata*** nei casi in cui sussistono dei pesi in corrispondenza delle differenti modalità. Come esempio, sviluppiamo i calcoli dell'esempio 4.6 del testo:

```
R==> studx=c(25,30,30,28,22,27,25,30,24,30,20,27,26,28,22,30,30,22);
R==> study=c(25,30,27,24,22,30,25,28,30,30,27,20,28,26,22,22,30,30);
R==> cred=c(4,6,4,5,12,8,9,9,7,5,10,5,10,6,10,4,4,12);

R==> mean(studx); mean(study);
[1] 26.444
[1] 26.444

R==> mediapx=sum(studx*cred)/sum(cred); mediapx;
[1] 25.538

R==> mediapy=sum(study*cred)/sum(cred); mediapy;
[1] 26.485
```

È confermato, per i dati di questo esempio, che la media aritmetica (coincidente per i due studenti) cambia in modo sensibile se viene calcolata ponderando i voti per i CFU.

R4.2 Moda, mediana e quantili

La ***moda*** è ben definita per qualsiasi variabile e per il suo calcolo occorre individuare a quale modalità corrisponde alla frequenza massima. L'esempio mostra come ottenere questo risultato nel caso specifico del calcolo della moda per la variabile `componenti` nel data set `studenti`, utilizzando la funzione `which.max(vettore)` che fornisce l'indice di `vettore` cui corrisponde il massimo:

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> componenti=dati[,10];
R==> freq=tabulate(componenti);
R==> which.max(freq);              ### moda
[1] 4
```

Quindi, la moda del numero di componenti è 4, cioè le famiglie più diffuse sono quelle con 4 componenti.

Si osservi che la procedura qui evidenziata può essere replicata correttamente nei casi in cui le modalità assumono valori da 1 al massimo, perché la funzione `tabulate` (in assenza di opzioni) calcola la distribuzione di frequenze in tale intervallo.

La ***mediana*** si calcola con la funzione `median()` che applica le convenzioni illustrate nel testo.

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> componenti=dati[,10];

R==> median(componenti);                    ### mediana
[1] 4
```

Si osservi come questi risultati si possono ottenere anche con le istruzioni seguenti:

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> with(dati, median(componenti));    ### mediana
[1] 4
R==> with(dati, mean(componenti));    ### media aritmetica
[1] 4.5125
```

Inoltre, anche la funzione `median()` consente di rimuovere valori mancanti prima del calcolo:

```
R==> esempio=c(1,3,3,3,7,4,5,2,1,NA,3, 2,4,4,5,6,NA,NA);
R==> median(esempio, na.rm=TRUE);
[1] 3
```

Infine, una nota di cautela: se il *data frame* è costituito da variabili tutte quantitative la funzione `mean` genera la media aritmetica delle variabili per ciascuna colonna; invece, la funzione `median` genera la mediana di tutti i valori presenti nel data frame (e quindi non ha generalmente alcun significato, e provoca un errore). In questo secondo caso, occorre usare la funzione `apply` di cui tratteremo nel paragrafo 6.5. L'esempio mostra cosa succede per il data set `studenti`:

```
R==>dati=read.table("studenti.txt", header=TRUE);
R==> mean(dati[,3:5]);   ### calcola la media delle variabili
     eta     peso altezza
 21.851   63.381 169.012

R==> median(dati[,3:5]); ### genera un errore
```

```
Errore in median.default(dati[, 3:5]) : necessario dati numerici

R==> apply(dati[,3:5],2,median); ### questo è il modo corretto
     eta    peso altezza
    20.8    60.0   168.0
```

Per i *quantili* di variabile, si usa la funzione quantile(variabile,probs=vettore) dove vettore indica le proporzioni al di sotto delle quali vanno individuate le modalità. Per esempio, le istruzioni per calcolare *quartili* e *decili* di variabile sono, rispettivamente:

```
R==> quantile(variabile,probs=c(0.25,0.50,0.75)); ### quartili
R==> quantile(variabile,probs=(1:9)/10);          ### decili
```

L'esempio mostra quartili e decili, rispettivamente, calcolati per la variabile componenti nel data set studenti.

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> componenti=dati[,10]

R==> quantile(componenti,probs=c(0.25,0.50,0.75));
25%  50%  75%
  4    4    5          ### 1°,2°,3° quartile

R==> quantile(componenti,probs=(1:9)/10);
10%  20%  30% 40% 50% 60% 70% 80% 90%
  3    4    4   4   4   5   5   5   6  ### 1°,2°,...,9° decile
```

Se si utilizza l'opzione names=FALSE la funzione non produce l'elenco delle percentuali. Anche tale funzione consente la rimozione di dati mancanti prima del calcolo.

La funzione fivenum(variabile) fornisce direttamente i 5 numeri che caratterizzano in modo sintetico alcuni importanti indici di posizione. In breve, essa calcola e fornisce ordinatamente minimo, primo quartile, mediana, terzo quartile, massimo di una variabile quantitativa.

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> componenti=dati[,10];
R==> fivenum(componenti);
[1]  1  4  4  5 12
```

Ricordiamo, ancora una volta, che i precedenti risultati possono essere ottenuti utilizzando la funzione with(), cioè con le istruzioni:

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> with(dati, quantile(componenti,probs=c(0.25,0.50,0.75)));
R==> with(dati, fivenum(componenti));
```

L'ambiente R utilizza un suo modo di interpolare le osservazioni per il calcolo dei quantili e che esistono molte opzioni per modificare tale criterio. Per questo, si possono osservare delle discrepanze numeriche se si calcolano i quantili con software differenti.

Infine, per variabili quantitative, R possiede la funzione summary() che esplicita minimo, massimo, media e mediana della variabile in questione.

R4.3 Altri indici di posizione

Per ottenere la *media troncata* (*trimmed*), si inserisce nella funzione mean() l'opzione trim=quota, dove quota è la proporzione (quindi, tra 0 e 1) delle modalità più piccole e più grandi da escludere prima di calcolare la media aritmetica.

Il *valore centrale* (*midrange*) di una serie di modalità, contenute nel vettore variabile si ottiene semplicemente facendo la semisomma del minimo e massimo di variabile, e quindi con l'istruzione: mean(range(variabile)).

Infine, l'esempio che ora presentiamo (e che contiene alcuni dati anomali) mostra come ottenere la ***media secondo Winsor*** sostituendo le prime due e le ultime due modalità ordinate con la terza e la terzultima modalità, rispettivamente. Si osservi che, per ottenere le medie secondo Winsor, è necessario ordinare preliminarmente i dati mediante la funzione sort(variabile).

```
R==> variabile=c(-10,0,0,2,2,3,4,-2,30,-1,0,5,3,8,50);
R==> mean(variabile);              ### media aritmetica
[1] 6.2667

R==> mean(variabile,trim=0.1);  ### media troncata al 10%
[1] 4.1538

R==> mean(range(variabile));    ### valore centrale
[1] 20

R==> varsort=sort(variabile);   ### ordina la variabile
R==> n=length(varsort);         ### lunghezza variabile

### sostituisce i 2+2 dati estremi
R==> varsort[1]=varsort[3];  varsort[2]=varsort[3];
R==> varsort[n]=varsort[n-2];varsort[n-1]=varsort[n];

R==> mean(varsort);                ### media secondo Winsor
[1] 2.6667
```

In questo esempio, la presenza di taluni valori anomali provoca una sostanziale alterazione della media aritmetica (e del valore centrale, ovviamente), che viene corretta dalla media troncata ed ancor più dalla media secondo Winsor.

Per calcolare la ***media geometrica*** (definita solo per una variabile quantitativa con valori tutti positivi) e la ***media armonica*** di una serie di modalità presentiamo il seguente esempio:

```
R==> variabile=c(1, 2, 2, 3, 4, 2, 1, 5, 3, 8, 5, 7);

R==> geom=exp(mean(log(variabile))); ### media geometrica
R==> harm=1/mean(1/variabile);       ### media armonica

R==> harm; geom; mean(variabile); ### le tre medie
[1] 2.3601
[1] 2.9317
[1] 3.5833
```

L'esempio conferma che – quando tutte le modalità sono positive – la media armonica è non superiore alla media geometrica che, a sua volta, è non superiore alla media aritmetica.

Capitolo quinto

Indici statistici di variabilità

5.1 Introduzione

In questo Capitolo esamineremo alcune misure della variabilità di una distribuzione di frequenza, distinguendo tra dispersione rispetto ad un centro, dispersione delle modalità ordinate, mutua variabilità e mutevolezza delle frequenze. In tale ambito, daremo un certo rilievo allo studio della concentrazione di una variabile trasferibile.

Per comprendere il ruolo centrale della variabilità per qualsiasi problematica affrontata dalla Statistica, esaminiamo un caso concreto che mostra come alla stessa posizione centrale possano corrispondere situazioni molto diverse in termini di variabilità.

Esempio 5.1

Le seguenti distribuzioni delle Età (in anni compiuti) di tre gruppi di laureati possiedono tutte media e mediana pari a 25 anni, ma è evidente una differente dispersione attorno a tale valore:

$$\begin{aligned} A:&\ \{22,22,23,23,24,25,26,27,27,28,28\} \\ B:&\ \{22,22,22,22,22,25,28,28,28,28,28\} \\ C:&\ \{25,25,25,25,25,25,25,25,25,25,25\} \end{aligned}$$

- Per il gruppo A, il valore 25 è una sintesi accettabile per la posizione della distribuzione delle età, anche se un solo laureato ha 25 anni.
- Per il gruppo B, la media 25 è il valore che si colloca al centro della distribuzione di frequenza, ma la distribuzione si bipartisce tra due soli valori: 22 e 28.
- Per il gruppo C, tutti i valori sono pari a 25 che risulta essere media, moda e mediana della distribuzione; in tal caso, la conoscenza della media riassume completamente l'intera distribuzione.

□

Quindi, come l'esempio ha confermato, la sola utilizzazione di indici di posizione non è sufficiente a discriminare tra situazioni tra loro molto differenziate ma potrebbe, anzi, in talune circostanze, produrre delle sintesi fuorvianti.

Sul piano concettuale, possiamo dire che *la variabilità di un fenomeno è la sua attitudine ad assumere differenti modalità*. Le differenti misure di tale attitudine devono rispettare alcuni principi generali richiesti dal concetto stesso di variabilità.

Tra i requisiti necessari per una misura di variabilità, risultano fondamentali ed intuitivi i seguenti:

i) ogni indice di variabilità è sempre maggiore o uguale a 0;

ii) ogni indice di variabilità calcolato per una distribuzione costante è 0;

iii) aggiungendo una costante ad una variabile, l'indice di variabilità non deve cambiare.

D'altra parte, la variabilità di una variabile X può essere misurata rispetto a più punti di vista:

- *variabilità* delle singole modalità $x_1, x_2, \ldots, x_n$ *rispetto ad un valore di posizione* (per esempio, la media, la mediana, etc.) mediante una sintesi degli scarti tra le singole modalità e il valore di riferimento;
- *variabilità delle modalità ordinate* $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ mediante indici derivati dalla funzione di ripartizione empirica;
- *variabilità reciproca* (mutua) tra tutte le modalità considerate a due a due, mediante una sintesi delle diversità esistenti tra tutte le coppie di modalità;
- *variabilità rispetto alle sole frequenze relative*, mediante una sintesi della mutevolezza delle frequenze.

Evidentemente, mentre le prime tipologie possono riferirsi solo a variabili quantitative (poiché occorre tenere conto nel calcolo dell'indice delle singole modalità), l'ultima può essere applicata anche a variabili qualitative (perché solo le frequenze sono coinvolte nel calcolo di tali indici).

5.2 Variabilità rispetto ad un centro

Un fenomeno può variare nel tempo, sul territorio, tra i soggetti e un indice di variabilità deve misurare questi aspetti. È ragionevole, allora, misurare la variabilità di un fenomeno controllando se le singole unità statistiche presentano modalità più o meno stabili rispetto ad un indice di posizione, il quale viene assunto come rappresentativo della intera distribuzione di frequenza.

Se – come spesso avviene – si assume che il punto di riferimento sia la media μ, allora si tratta di sintetizzare le differenze tra le singole modalità rispetto a μ, cioè le quantità:

$$x_1 - \mu,\ x_2 - \mu,\ \ldots,\ x_n - \mu\,,$$

che si definiscono *scarti dalla media*, o semplicemente ***scarti***.

o Riprendendo i dati dell'esempio 5.1 si vede subito che, essendo in ogni caso la media pari a $\mu = 25$, per il gruppo B tali differenze sono più elevate in valore assoluto rispetto alle analoghe differenze del gruppo A; invece, tali differenze sono tutte identicamente nulle per il gruppo C. Quindi, utilizzare gli scarti per graduare la variabilità può essere un buon criterio.

D'altra parte, per quanto dimostrato nel paragrafo 4.4, la somma algebrica degli scarti è nulla (proprietà ii) mentre è minima la somma dei quadrati degli scarti (proprietà iv). Quindi, è opportuno sintetizzare la variabilità di una variabile X tramite i quadrati degli scarti: perché i quadrati sono sempre maggiori o uguali a 0 e quindi non si compensano tra di loro, ed anche perché la loro somma costituisce un minimo rispetto ad altre proposte.

L'indice più importante per misurare la variabilità di una distribuzione è espresso dalla media degli scarti al quadrato, si chiama ***varianza*** ed è universalmente indicato con il simbolo σ^2:

$$\sigma^2 = \frac{(x_1-\mu)^2 + (x_2-\mu)^2 + \ldots + (x_n-\mu)^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\,.$$

Se i dati sono disponibili mediante una seriazione la precedente formula si modifica per tenere conto dei raggruppamenti delle modalità:

$$\sigma^2 = \frac{(x_1-\mu)^2 n_1 + (x_2-\mu)^2 n_2 + \ldots + (x_k-\mu)^2 n_k}{n_1 + n_2 + \ldots + n_k} = \frac{1}{n}\sum_{i=1}^{k}(x_i-\mu)^2 n_i\,.$$

È agevole convincersi che *la varianza è una misura sempre non-negativa* (perché media di quadrati).

Essa varia da un minimo di 0 (se tutte le modalità coincidono tra di loro: in tal caso, tutte le unità presentano lo stesso valore costante ed uguale a μ) ad un massimo che può crescere indefinitamente (perché gli scarti possono anche essere estremamente lontani dalla media aritmetica).

Se si suppone che la media μ sia predeterminata, la situazione di massima variabilità si presenta quando una sola modalità racchiude l'intero fenomeno e le rimanenti $n-1$ valgono 0. In tal caso, per semplice sostituzione nella formula della varianza, risulta che: $\sigma^2 \leq \mu^2(n-1)$.

o In alcune situazioni, è utile conoscere che la varianza non può superare la quantità var_{max} definita nello schema seguente:

$$var_{max} = \left(\frac{max(x_i)-min(x_i)}{2}\right)^2, \quad \text{in generale,}$$

$$var_{max} = \left(\frac{max(x_i)-min(x_i)}{3}\right)^2, \quad \text{se la distribuzione è unimodale;}$$

Il vincolo $0 \leq \sigma^2 \leq var_{max}$ può risultare particolarmente utile nei casi in cui non si possiede alcuna informazione sulla variabile X tranne i suoi estremi. Per esempio, se si conosce che la variabile X è unimodale e può in ogni caso assumere valori compresi nell'intervallo (140, 200), allora: $\left(\frac{200-140}{3}\right)^2 = 400$, per cui: $0 \leq \sigma^2 \leq 400$.

o Se occorre svolgere il calcolo della varianza, è consigliabile utilizzare una semplificazione che deriva dallo sviluppo algebrico della sua definizione, secondo i passaggi seguenti:

$$\begin{aligned}\sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i)^2 - 2\mu\frac{1}{n}\sum_{i=1}^{n}(x_i) + \frac{1}{n}\sum_{i=1}^{n}(\mu)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i)^2 - 2\mu(\mu) + (\mu)^2\end{aligned}$$

e, quindi, alla fine, risulta:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i)^2 - (\mu)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i)^2 - \left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)^2.$$

Nel caso di una seriazione, l'ultima espressione diventa:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i^2\, n_i) - \left(\frac{1}{n}\sum_{i=1}^{k}x_i\, n_i\right)^2.$$

In ogni caso, si può affermare che *la varianza si calcola sottraendo dalla media aritmetica dei quadrati il quadrato della media aritmetica.* Quest'ultima formulazione è particolarmente agevole perché richiede, in effetti, il solo calcolo della somma delle modalità e delle modalità al quadrato.

Riassumiamo, quindi, le principali caratteristiche e *proprietà della varianza*:

1. La varianza è sempre un numero non-negativo (cioè maggiore o uguale a 0).
2. La varianza è 0 se e solo se la variabile X è una costante, cioè assume in tutte le unità statistiche il medesimo valore. In tal caso, la variabile X si chiama *degenere*.
3. Se alla variabile X si aggiunge una qualsiasi costante finita c, la sua varianza non cambia.
4. Se la variabile X si moltiplica per una qualsiasi costante finita b, la sua varianza viene moltiplicata per b^2.

Esempio 5.2

La variabile X = «Numero di componenti della famiglia» in un certo collettivo di $n = 5$ unità statistiche ha presentato i seguenti valori {3, 1, 2, 4, 2}.

Quindi, la media aritmetica è: $\mu = 2.4$ e gli scarti (dalla media aritmetica) saranno: $\{3-2.4,\ 1-2.4,\ 2-2.4,\ 4-2.4,\ 2-2.4\}$ cioè $\{0.6, -1.4, -0.4, 1.6, -0.4\}$.

Trattandosi di poche osservazioni, i calcoli per ottenere la varianza si possono organizzare come nella tabella seguente:

Unità i	Modalità (x_i)	(Modalità)^2 $(x_i)^2$	Scarti dalla media $(x_i-\mu)$	(Scarti dalla media)^2 $(x_i-\mu)^2$
1	3	9	+0.6	0.36
2	1	1	−1.4	1.96
3	2	4	−0.4	0.16
4	4	16	+1.6	2.56
5	2	4	−0.4	0.16
Somme	12	34	0.0	5.20

Applicando la definizione della varianza si ha:

$$\sigma^2 = \frac{1}{5}\sum_{i=1}^{5}(x_i-2.4)^2 = \frac{5.20}{5} = 1.04.$$

Questo risultato poteva anche dedursi utilizzando la precedente formula alternativa per il calcolo della varianza. Infatti, si ha:

$$\sigma^2 = \frac{1}{5}\sum_{i=1}^{5}(x_i)^2 - \left(\frac{1}{5}\sum_{i=1}^{5}x_i\right)^2 = \frac{34}{5} - \left(\frac{12}{5}\right)^2 = 6.80 - 5.76 = 1.04.$$

□

Una difficoltà nella interpretazione della varianza deriva dal fatto che essa è espressa nell'unità di misura della variabile al quadrato. Per questo, si introduce lo ***scarto quadratico medio*** (chiamato, talvolta, deviazione standard, dall'inglese: *standard deviation*) che è definito come la radice quadrata della varianza, ed è ovviamente indicato con σ:

$$\sigma = \sqrt{\frac{(x_1-\mu)^2 + (x_2-\mu)^2 + \ldots + (x_n-\mu)^2}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2}\ .$$

Sono ovvie le modifiche necessarie per pervenire ad analoghe espressioni quando i dati sono disponibili mediante una seriazione.

D'altra parte, poiché sia la varianza che lo scarto quadratico medio dipendono dall'unità di misura (infatti, sono indici assoluti) è opportuno considerare anche indici relativi (che, invece, sono svincolati dall'unità di misura della variabile).

Un indice relativo molto usato è ottenuto facendo il rapporto tra lo scarto quadratico medio σ e la media aritmetica μ. Esso si chiama ***coefficiente di variazione*** e sarà indicato nel seguito con il simbolo CV:

$$CV = \frac{\sigma}{\mu} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2}}{\mu} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\mu}\right)^2}.$$

- Il coefficiente di variazione – che spesso viene presentato in termini percentuali, moltiplicando il valore di CV per 100 – è indipendente dall'unità di misura, cioè è un numero puro. Esso misura la variazione media del fenomeno in rapporto alla sua media aritmetica ed è utile per confrontare la variabilità relativa di un fenomeno in circostanze differenti: la variabilità della distribuzione per Età tra le varie regioni, la distribuzione dei Redditi per nazioni e per anno, la variabilità del Peso rispetto al sesso, etc.
- Inoltre, il coefficiente di variazione è necessario tutte le volte che si intende confrontare la variabilità di due fenomeni espressi in unità di misure non confrontabili. Per esempio, la variabilità del Peso rispetto a quella dell'Altezza; la variabilità dei Consumi di carburante rispetto alla variabilità dell'Usura dei pneumatici per una data marca di autovetture, etc.
- Tra i difetti di CV va segnalato il fatto che tale indice ha significato solo se $\mu > 0$ ed inoltre tende a diventare infinitamente grande (e quindi a non essere utile per confrontare la variabilità di fenomeni differenti) se $\mu \to 0$.

Fra le numerose altre proposte di indici di variabilità rispetto ad un centro, ci limitiamo a citare lo *scostamento semplice dalla media* $S(\mu)$ e lo *scostamento semplice dalla mediana* $S(Me)$, per i quali le differenze dal centro di riferimento (la media e la mediana, rispettivamente) sono effettuate in valore assoluto.

Essi sono definiti dalle seguenti espressioni:

$$S(\mu) = \frac{1}{n}\sum_{i=1}^{n}|x_i-\mu|; \qquad S(Me) = \frac{1}{n}\sum_{i=1}^{n}|x_i-Me|.$$

Per essi, si dimostra la disuguaglianza:

$$S(Me) \le S(\mu) \le \sigma.$$

Esempio 5.3

Riprendiamo i dati dell'esempio 5.2 per calcolare i successivi indici di variabilità proposti. In primo luogo, lo scarto quadratico medio e il coefficiente di variazione sono:

$$\sigma = \sqrt{1.04} = 1.01980; \qquad CV = \frac{1.01980}{2.4} = 0.42492.$$

Il vantaggio di σ rispetto a σ^2 risiede nel fatto che lo scarto quadratico medio è espresso nella stessa unità di misura di μ (e, quindi della variabile originaria). Invece, dal coefficiente di variazione risulta che la variabilità è una frazione 0.42 della media, cioè il 42% della media.

Ordinando i dati, si ottiene la mediana: $Me = 2$ e, ricordando che la media è $\mu = 2.4$, si trae la tabella seguente:

Unità i	Modalità (x_i)	| Scarti dalla media | $\lvert x_i-\mu\rvert$	| Scarti dalla mediana | $\lvert x_i-Me\rvert$
1	3	0.6	1
2	1	1.4	1
3	2	0.4	0
4	4	1.6	2
5	2	0.4	0
Somme	12	4.4	4

Pertanto, lo scostamento semplice medio e lo scostamento semplice mediano sono rispettivamente pari a:

$$S(\mu) = \frac{1}{5}\sum_{i=1}^{5}|x_i-2.4| = \frac{4.4}{5} = 0.88$$

$$S(Me) = \frac{1}{5}\sum_{i=1}^{5}|x_i-2| = \frac{4}{5} = 0.80$$

Questi risultati confermano la disuguaglianza: $S(Me) \le S(\mu) \le \sigma$, che nel caso specifico corrisponde a: $0.80 < 0.88 < 1.01980$. □

5.3 Variabilità delle modalità ordinate

Alcune misure di variabilità possono essere derivate dalla *funzione di ripartizione empirica* (paragrafo 3.5), e quindi dalle modalità ordinate; in particolare, tali misure utilizzano i quantili della distribuzione di frequenza.

Infatti, se la funzione di ripartizione è molto ripida le modalità della variabile sono tutte assai vicine al valore di posizione (media, mediana, etc.); invece, se tale funzione tarda a raggiungere il valore 1, allora vi è una grande variabilità nelle osservazioni. Per questo, il confronto tra i quantili costituisce un approccio idoneo per costruire altri indici di variabilità.

- Il ***campo di variazione*** (=*range*) $R(X)$ è la differenza tra il massimo e il minimo delle modalità della popolazione, per cui:

$$R(X) = \max(x_i) - \min(x_i).$$

Esso è influenzato anche da un solo valore atipico (perché molto basso o molto alto), il che rende l'indice vulnerabile ad errori e/o situazioni eccezionali. È, quindi, preferibile in generale esaminare la variazione tra quantili intermedi rispetto a quelli estremi.

• Il ***campo di variazione interquartile*** (=*Inter-Quartile Range*) $IQR(X)$ è definito come la differenza tra il terzo e il primo quartile, cioè:

$$IQR(X) = Q_3 - Q_1.$$

Talvolta, l'indice IQR si divide per 2, oppure – per renderlo confrontabile con lo scarto quadratico medio di una particolare distribuzione teorica (di cui parleremo diffusamente nel paragrafo 10.3) – si utilizza come misura di variabilità l'indice: $0.7413 \times IQR$.

L'indice IQR è basato sulle informazioni espresse solo dal 50% della popolazione collocata al centro della distribuzione; per questo, è anche opportuno considerare indici fondati su una maggiore porzione della popolazione.

• Il ***campo di variazione interdecile*** $IDR(X)$ (=*Inter-Decile Range*) è definito come la differenza tra il nono e il primo decile, cioè:

$$IDR(X) = D_9 - D_1.$$

Tale quantità è sufficientemente resistente perché non è influenzata dal 20% della popolazione che presenta modalità estreme, e tiene conto della variabilità espressa dall'80% della distribuzione di frequenza. Per motivi analoghi ai precedenti, talvolta, si utilizza come indice di variabilità la quantità: $0.3917 \times IDR$.

Esempio 5.4

Nella successiva tabella presentiamo alcuni indicatori di posizione e di variabilità per la variabile Età nella popolazione degli universitari componenti il data set STUDENTI, distinta rispetto al sesso.

Indicatori	Maschi	Femmine	Totale
n	1079	1680	2759
$min(x_i)$	19.1	19.1	19.1
D_1	19.5	19.3	19.3
Q_1	20.4	19.8	20.2
Me	21.4	20.7	20.8
Q_3	23.7	21.7	22.3
D_9	26.6	23.7	24.9
$max(x_i)$	63.5	56.1	63.5
μ	22.770	21.260	21.851
σ	4.433	2.678	3.549
CV	0.195	0.126	0.162
$R(X)$	44.4	37.0	44.4
$IDR(X)$	7.10	4.40	5.60
$IQR(X)$	3.30	1.90	2.10

Questi indici costituiscono una informazione piuttosto completa per valutare posizione e variabilità della distribuzione delle Età rispetto alla mutabile individuata (il sesso dello studente). Peraltro, trattandosi di una popolazione numerosa e riferita a dati particolarmente accurati, le risultanze che se ne derivano possono essere estese a collettivi simili.

Per il complesso della popolazione, l'Età media supera di 1 anno l'Età mediana indicando che esistono studenti/studentesse particolarmente anziani rispetto al complesso della distribuzione. Ciò deriva da un fatto strutturale: mentre esiste un limite inferiore per la frequenza al 1° anno di un corso universitario (almeno 18 anni al momento della rilevazione) non vi è teoricamente limite superiore per la partecipazione ai corsi (infatti, uno studente ha dichiarato oltre 63 anni e una studentessa oltre 56 anni). Ovviamente, la popolazione esaminata è in grandissima parte giovanile perché il 90% degli studenti ha meno di 24.9 anni (come emerge dal valore del 9° decile D_9).

I dati mostrano anche una sensibile differenza di Età rispetto al Sesso: le studentesse sono sistematicamente più giovani dei loro colleghi, come confermato da tutti gli indici di posizione. La differenza tra le mediane è di appena 0.7 anni (8 mesi) ma la differenza tra le medie sale a 1.5 anni (18 mesi) per giungere quasi a 3 anni di differenza se si confronta il 9° decile. Quindi, la distribuzione delle Età degli studenti è spostata in modo abbastanza netto verso destra rispetto all'analoga distribuzione delle Età delle studentesse.

Infine, tutti gli indicatori di variabilità per le studentesse sono più bassi rispetto a quelli per gli studenti: ciò implica che le risultanze sulla posizione (cioè quanto detto circa la minore Età media delle studentesse) sono molto affidabili e più nette per le studentesse, perché la loro distribuzione delle Età è meno variabile di quella maschile.

□

5.4 Mutua variabilità

Numerosi ed importanti problemi reali (come la disuguaglianza dei redditi, la ripartizione delle risorse naturali, economiche e finanziarie tra individui, gruppi, nazioni) richiedono di misurare la diversità tra le singole unità statistiche, piuttosto che rispetto ad un ipotetico centro rappresentativo.

Oggetto di questo paragrafo e del successivo è la mutua variabilità di una ***variabile trasferibile***. Infatti, solo la trasferibilità di un bene, servizio, territorio, evento, etc. rende possibile la determinazione teorica della variabilità minima e massima all'interno di un ammontare complessivo prefissato. In sostanza, una variabile quantitativa è trasferibile se è possibile ipotizzare che l'ammontare complessivo del fenomeno registrato su una prefissata popolazione possa essere distribuito tra le unità statistiche senza modificarne l'ammontare complessivo.

○ Sono *variabili trasferibili* il reddito, il consumo di un bene, i finanziamenti alle imprese, la scelta degli investimenti nei settori produttivi, la distribuzione della popolazione e delle risorse economiche, ambientali, energetiche tra gli Enti locali (comuni, province, regioni, comprensori, aziende sanitarie) o tra le nazioni e i continenti della Terra, la superficie agricola utilizzata (SAU) rispetto a quella delle aziende agrarie che la possiedono, gli studenti universitari rispetto ai corsi di laurea e di diploma prescelti, le nascite rispetto al comune di residenza della madre, le abitazioni in rapporto al territorio, la piovosità annua in un determinato bacino rispetto ai sotto-periodi di rilevazione (mesi, giorni, stagioni), le molecole di una sostanza in un ambiente fluido ben definito, gli arrivi dei clienti negli esercizi alberghieri rispetto al mese dell'anno, il numero di pagine dei capitoli di un libro rispetto al totale delle pagine di cui è composto, e così via.

Per ovvie ragioni, sono *variabili non-trasferibili* l'età, il peso, l'altezza, la velocità delle auto, il numero di componenti di una famiglia, etc.

Per la discussione successiva è fondamentale definire con cura le due situazioni estreme, di minima e di massima mutua variabilità, in quanto tutte le situazioni intermedie saranno poi riferite a tali estremi. Lo faremo fissando l'ammontare complessivo $\sum x_i$ di una variabile trasferibile X, e di conseguenza la sua media: $\mu = \frac{1}{n}\sum x_i$.

- La *variabilità minima* si verifica quando non c'è la benché minima diversità tra l'ammontare della variabile trasferibile nelle differenti unità statistiche. In tal caso, si verifica:

$$x_1 = x_2 = \ldots = x_n = \mu.$$

In tale situazione, qualunque misura di mutua variabilità deve essere nulla.

- La *variabilità massima* si verifica se una sola unità statistica racchiude in sé l'intero ammontare $\sum x_i = n\,\mu$ della variabile trasferibile, lasciando alle altre $(n-1)$ unità un ammontare pari al valore 0, cioè quando:

$$x_1 = \sum x_i = n\,\mu; \quad x_2 = x_3 = \ldots = x_n = 0.$$

In tale situazione, qualunque misura di mutua variabilità deve assumere il valore massimo, non essendo possibile una disuguaglianza reciproca superiore a questa.

- Da quanto detto, segue che una misura della mutua variabilità si ottiene esaminando tutte le differenze tra le modalità a due a due $|x_i - x_j|$ per $(i \neq j) = 1, 2, \ldots n$, facendone poi una sintesi tramite una opportuna media.
- È necessario considerare il valore assoluto delle differenze per evitare che, nel fare la sintesi, ogni confronto $(x_i - x_j)$ si annulli con il confronto opposto: $(x_j - x_i)$.
- Poiché tutti i possibili confronti tra le n unità statistiche sono pari a $n(n-1)$, avendo escluso i confronti tra una unità e sé stessa, quando i dati sono disponibili in serie $(x_1, x_2, \ldots, x_n)$, la sintesi più immediata consiste nella media aritmetica di tali differenze prese in valore assoluto.

Allora, la ***differenza semplice media*** per dati disponibili in serie è definita da:

$$\Delta = \frac{\sum_{i \neq j=1}^{n} |x_i - x_j|}{n\,(n-1)}.$$

- Nella precedente formula, il simbolo $\sum_{i \neq j=1}^{n}$ indica che occorre sommare tutte le differenze in valore assoluto facendo cambiare valore sia ad i che a j (per tutti gli interi compresi tra 1 e n) ma in modo che tali indici siano tra loro differenti.

- Se, invece, i dati sono disponibili mediante una seriazione $(x_1, x_2, \ldots, x_k)$ con rispettive frequenze assolute $(n_1, n_2, \ldots, n_k)$ tali che $\sum n_i = n$, allora tutte le possibili differenze del tipo $|x_i - x_j|$ saranno in numero pari a $n_i\,n_j$. Pertanto, il calcolo della differenza semplice media diventa:

$$\Delta = \frac{\sum_{i \neq j=1}^{k} |x_i - x_j|\,n_i\,n_j}{n\,(n-1)}.$$

- Il *minimo di* Δ vale 0 e si verifica quando i dati coincidono tra di loro, per cui tutte le differenze sono nulle (è la situazione che prima abbiamo definito di minima mutua variabilità).
- Il *massimo di* Δ si verifica quando $(n-1)$ unità statistiche possiedono 0 ed una sola unità possiede l'ammontare complessivo della variabile $\sum x_i$, che è uguale a: $n\,\mu$. In tal caso, sostituendo $x_1 = n\,\mu$ e $x_2 = x_3 = \cdots = x_n = 0$ (per esempio) nella definizione di Δ, dopo qualche agevole passaggio algebrico, si ottiene: $\Delta = 2\,\mu$.

Pertanto, per qualsiasi variabile trasferibile X, la differenza semplice media è sempre compresa tra 0 e 2μ, cioè:

$$0 \leq \Delta \leq 2\mu\ .$$

Dividendo ambo i membri di tale disuguaglianza per 2μ si deduce che:

$$0 \leq \frac{\Delta}{2\mu} \leq 1.$$

Quindi, l'indice $R = \dfrac{\Delta}{2\,\mu}$, definito come il rapporto tra la differenza semplice media e il doppio della media aritmetica, costituisce una misura *normalizzata*, compresa tra 0 e 1, della mutua variabilità di una variabile trasferibile, perché: $0 \leq R \leq 1$.

Tale indice è il ***rapporto di concentrazione di Gini*** che è quindi così definito:

$$R = \frac{\Delta}{2\,\mu}.$$

Nel caso di una serie di osservazioni ordinate $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, il rapporto di concentrazione si può esprimere nel modo seguente, che è di calcolo più semplice:

$$R = 2\,\frac{\sum_{i=1}^{n} i\,x_{(i)}}{(n-1)\sum_{i=1}^{n} x_{(i)}} - \frac{n+1}{n-1}.$$

Esempio 5.5

Il reddito mensile (espresso in migliaia di Euro) di una popolazione di 5 unità statistiche è espresso dalle seguenti modalità: {1, 3, 5, 7, 9}. Trattandosi di un esempio con poche osservazioni, mostriamo il calcolo del rapporto di concentrazione sia utilizzando direttamente la definizione (e, quindi, tramite la differenza semplice media) che mediante la formula più semplice (che richiede i dati già ordinati).

Per utilizzare la definizione di Δ occorre calcolare la somma di tutte le differenze a due a due tra le modalità, e poi farne la media aritmetica, dividendo tale somma per il numero dei confronti, cioè:

$$\Delta = \frac{|1-3|+|1-5|+|1-7|+|1-9|+|3-1|+|3-7|+\ldots+|9-5|+|9-7|}{5\times 4}$$
$$= \frac{80}{20} = 4.$$

Quindi, essendo la media dei pari a: $\mu = 25/5 = 5$, si ottiene: $R = 4/(2\times 5) = 0.4$.

In alternativa, essendo i dati già ordinati, si ha:

$$\sum x_i = \sum x_{(i)} = 25; \quad \sum i\, x_{(i)} = 1\times 1 + 2\times 3 + 3\times 5 + 4\times 7 + 5\times 9 = 95;$$

per cui, applicando la formula semplificata per R si ha:

$$R = 2\times \frac{95}{(5-1)\times 25} - \frac{5+1}{5-1} = 0.4\,.$$

Il collettivo esaminato presenta, quindi, una mutua variabilità espressa dall'indice $R = 0.4$, pari al 40% di quella massima ipotizzabile.

□

In questo paragrafo, l'indice R è stato derivato mediante una normalizzazione della differenza semplice media. Esso, tuttavia, ha un significato proprio di particolare rilevanza per fenomeni economici, demografici, ambientali e sociali, come sarà esaminato in dettaglio nel paragrafo seguente.

5.5 Concetto e misura della concentrazione

La concentrazione di una variabile X deriva dalla possibilità di «trasferire» l'ammontare del fenomeno da una unità statistica ad un'altra, avvicinandosi o allontanandosi dalla situazione di equidistribuzione dell'ammontare complessivo della variabile.

o La comprensione di tale concetto riesce più agevole, e persino visibile, supponendo – a scopo esemplificativo – di dover discutere della *concentrazione dei redditi* tra un gruppo di unità statistiche ben definite (individui, famiglie, gruppi sociali), anche se il concetto e la misura della concentrazione si possono estendere a tutte le variabili trasferibili. Così, diremo che *il reddito di un paese è tanto più concentrato quanto più il reddito complessivo è posseduto da una frazione modesta delle unità statistiche, ovvero quanto più poveri vi sono in quel paese.*

Quindi, la concentrazione cresce con il crescere della frazione di unità statistiche che possiede la variabile in misura inferiore alla media ovvero con il diminuire delle unità statistiche che possiedono la variabile in misura superiore alla media.

Dalle affermazioni precedenti deriva che qualsiasi misura di concentrazione deve essere «sensibile» ai trasferimenti della variabile tra le unità statistiche nel senso che *ceteris paribus* se la modalità x_i cresce di una quantità $\delta > 0$ e contemporaneamente un'altra modalità x_j diminuisce dello stesso ammontare δ (per cui non si modifica né l'ammontare complessivo né la media della variabile nel collettivo), allora la concentrazione deve aumentare.

Come per la mutua variabilità, la concentrazione reale varia tra due estremi:

- si ha *concentrazione minima* quando l'ammontare complessivo della variabile è ripartito in misura uguale tra tutte le unità che, quindi, presentano modalità pari alla media;
- si ha *concentrazione massima* quando l'ammontare complessivo della variabile è posseduto da una sola unità statistica che lo possiede, quindi, in misura pari a $n\,\mu$, mentre le rimanenti $(n-1)$ unità posseggono 0.

Nei casi reali, si presenteranno situazioni intermedie nelle quali alcune unità statistiche possiedono la variabile in misura superiore alla media ed altre in misura inferiore. Obiettivo delle misure di concentrazione è proprio quello di valutare le situazioni reali in rapporto a questi due casi limite.

Ordiniamo, quindi, le modalità della variabile trasferibile X in senso non-decrescente: $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$, e indichiamo con:

- $p_i = \dfrac{i}{n}$, la frazione cumulata dei primi i redditieri, $i = 1, 2, \ldots, n$;
- $q_i = \dfrac{\sum_{j=1}^{i} x_{(j)}}{\sum_{j=1}^{n} x_{(j)}} = \dfrac{1}{n\,\mu}\sum_{j=1}^{i} x_{(j)}$, la frazione cumulata del reddito posseduto dai primi i redditieri, $i = 1, 2, \ldots, n$.

Ovviamente, p_i è la funzione di ripartizione empirica delle unità statistiche che assegna peso $\frac{1}{n}$ a ciascuna di esse, mentre q_i è la funzione cumulata della frazione di reddito posseduto dai primi i individui.

Una rappresentazione grafica molto efficace si ottiene ponendo in ascissa i valori p_i e in ordinata i corrispondenti valori q_i $i = 0, 1, \ldots, n$ convenendo che: $p_0 = q_0 = 0$. Unendo tali punti si ottiene la spezzata di concentrazione, nota come ***curva di Lorenz.***

o Per una migliore comprensione di tale rappresentazione, discutiamo l'interpretazione dei punti della curva di Lorenz.

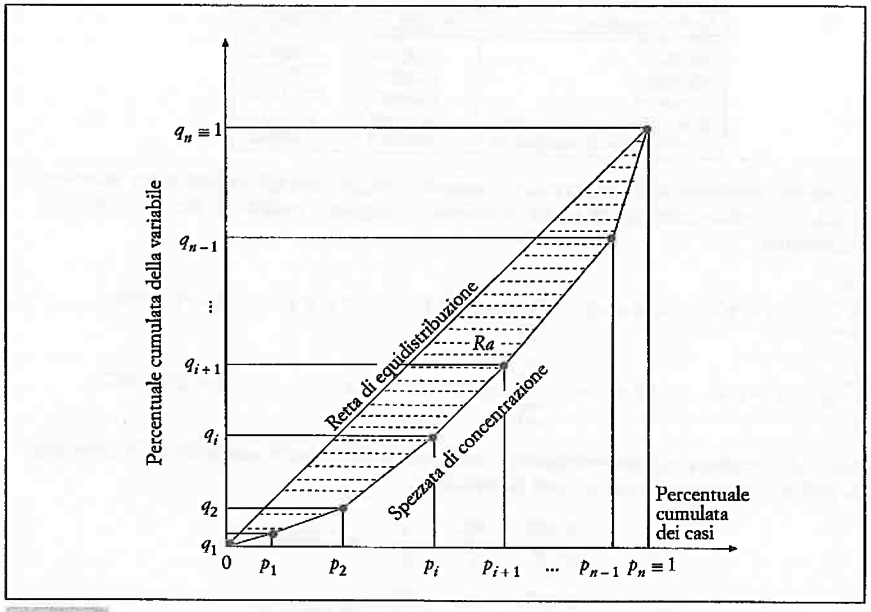

FIGURA 5.1. Curva di Lorenz e area di concentrazione.

- Il punto individuato dalla coppia di valori $(p_1, q_1) = \left(\frac{1}{n}, \frac{x_{(1)}}{\sum_{j=1}^{n} x_{(j)}}\right)$ rappresenta l'individuo più povero (il primo nella graduatoria) che è $\frac{1}{n}$ del totale degli individui e possiede un reddito $x_{(1)}$ che corrisponde alla frazione $\frac{x_{(1)}}{n\mu}$ del reddito complessivo $\sum_{j=1}^{n} x_{(j)} = n\mu$.

- Il punto individuato dalla coppia di valori $(p_2, q_2) = \left(\frac{2}{n}, \frac{x_{(1)} + x_{(2)}}{\sum_{j=1}^{n} x_{(j)}}\right)$ rappresenta i primi due individui più poveri (il primo e il secondo nella graduatoria) che assieme sono $\frac{2}{n}$ del totale degli individui e possiedono assieme un reddito $x_{(1)} + x_{(2)}$ il quale corrisponde alla frazione $\frac{x_{(1)} + x_{(2)}}{n\mu}$ del reddito complessivo $\sum_{j=1}^{n} x_{(j)} = n\mu$; e così via.

- Per un certo $i = 1, 2, \ldots, n$, il punto individuato dalla coppia di valori $(p_i, q_i) = \left(\frac{i}{n}, \frac{x_{(1)} + x_{(2)} + \cdots + x_{(i)}}{n\mu}\right)$ rappresenta i primi i individui più poveri (il primo, il secondo, ..., l'i-esimo nella graduatoria) che assieme sono $\frac{i}{n}$ del totale degli individui e possiedono assieme un reddito $x_{(1)} + x_{(2)} + \cdots + x_{(i)}$ il quale corrisponde alla frazione $\frac{x_{(1)} + x_{(2)} + \cdots + x_{(i)}}{n\mu}$ del reddito complessivo $\sum_{j=1}^{n} x_{(j)} = n\mu$; e così via.

- Infine, l'n-esimo punto individuato dalla coppia di valori:

$$(p_n, q_n) = \left(\frac{n}{n}, \frac{x_{(1)} + x_{(2)} + \cdots + x_{(n)}}{\sum_{j=1}^{n} x_{(j)}}\right) \equiv (1, 1)$$

rappresenta tutte le unità statistiche: infatti, la quota degli individui è $\frac{n}{n} = 1$ e, similmente, per la frazione del reddito complessivo da essi posseduto.

Evidentemente, $p_i \geq q_i$, $i = 0, 1, 2, \ldots, n$ perché, avendo ordinato i dati dal più povero al più ricco, il primo 10% (per esempio) delle unità può possedere al più il 10% (per esempio) del reddito complessivo: se così non fosse, esse non sarebbero le unità più povere.

D'altro canto, se fosse $p_i = q_i$, $i = 0, 1, 2, \ldots, n$, si avrebbe una situazione di prefetta equidistribuzione, perché il 10%, il 20%,..., il 90%, il 100% delle unità a partire da quelle più povere possiederebbero in tal caso, rispettivamente, il 10%, il 20%,..., il 90%, il 100% del reddito complessivo della popolazione.

Quindi, *le differenze* $(p_i - q_i) \geq 0, i = 0, 1, 2, \ldots, n$ *sono misure dirette della concentrazione* della variabile, perché la concentrazione aumenta in modo diretto con il valore di tali differenze.

o In effetti, la differenza $(p_i - q_i)$ misura in proporzione la quota del reddito complessivo che manca ai primi i individui per trovarsi in una posizione di equidistribuzione; più tale differenza è grande più i rimanenti $n - i$ individui concentrano su di essi una porzione rilevante dell'ammontare complessivo. Pertanto, è corretto ricercare una misura diretta della concentrazione della variabile trasferibile mediante una sintesi di tali differenze.

Nel 1914, lo statistico italiano Corrado Gini, riprendendo alcune intuizioni del 1912, propose come misura della diseguaglianza dei redditi la media aritmetica di tali differenze normalizzate (cioè delle quantità: $\frac{p_i - q_i}{p_i}, i = 1, 2, \ldots, n-1$), ponderate con pesi p_i, cioè mediante l'espressione:

$$\frac{\sum_{i=1}^{n-1} \left(\frac{p_i - q_i}{p_i}\right) p_i}{\sum_{i=1}^{n-1} p_i}.$$

Semplificando, si giunge all'indice R, universalmente noto come il ***rapporto di concentrazione di Gini*** e definito da:

$$R = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} p_i} = 1 - \frac{2}{n-1}\sum_{i=1}^{n-1} q_i .$$

- Nell'ultimo passaggio abbiamo sfruttato il fatto che:

$$\sum_{i=1}^{n-1} p_i = \sum_{i=1}^{n-1}\left(\frac{i}{n}\right) = \frac{1}{n}\sum_{i=1}^{n-1} i = \frac{1}{n}\frac{(n-1)(n)}{2} = \frac{n-1}{2}.$$

- Si dimostra che *il rapporto di concentrazione di Gini coincide con la differenza semplice media normalizzata rispetto al suo massimo*, introdotta nel paragrafo precedente. In altri termini, si dimostra che l'indice R è esattamente il rapporto tra Δ e 2μ.

- Se si dispone dei dati in serie $x_i, i = 1, 2, \ldots, n$ e di un algoritmo efficiente di ordinamento, pervenendo a $x_{(i)}, i = 1, 2, \ldots, n$, allora il modo più agevole per calcolare l'indice R è tramite la formula semplificata già esplicitata alla fine del paragrafo precedente.
- Il rapporto di concentrazione del Gini ha una importante interpretazione geometrica derivabile dalla Figura 5.1. Infatti, la *spezzata di concentrazione* è ottenuta congiungendo tra loro i punti $(p_i, q_i), i = 0, 1, \ldots, n$. Così, la spezzata di concentrazione si trova al di sotto della *retta di equidistribuzione*, che è la bisettrice degli assi e che corrisponde alla situazione di minima concentrazione possibile.

Nel caso di equidistribuzione della variabile (concentrazione minima), la spezzata coincide con la bisettrice degli assi positivi ovvero con la diagonale del quadrato unitario; all'opposto, nel caso di accentramento della variabile in un solo individuo (concentrazione massima), la spezzata è praticamente coincidente con i due lati, in basso e a destra, del quadrato unitario. Nei casi intermedi, più la spezzata è «tesa» verso la bisettrice degli assi positivi più siamo in presenza di una concentrazione bassa; viceversa, più è «adagiata» verso l'asse delle ascisse più ci si avvicina ad una situazione di elevata concentrazione.

Quindi, l'*area di concentrazione* – definita come quella compresa tra la spezzata di concentrazione e la retta di equidistribuzione – è una misura diretta della concentrazione.

Si dimostra che *il rapporto tra la misura dell'area della concentrazione ed il suo massimo (=0.5) coincide esattamente con il rapporto di concentrazione R.* Tale risultato consente di confrontare geometricamente la distribuzione della variabile trasferibile X in situazioni differenti (tempi, luoghi, circostanze).

Esempio 5.6

Supponiamo che il reddito in migliaia di Euro posseduto da due gruppi di persone, di numerosità $n = 9$, sia quello espresso dalle due distribuzioni seguenti:

$$X = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}; \qquad Y = \{1, 1, 1, 2, 2, 3, 3, 4, 28\}$$

È immediato convincersi che pur avendo i due gruppi lo stesso reddito medio, il secondo gruppo è notevolmente più variabile del primo, come è confermato dagli indici elencati nella tabella seguente:

Indici statistici	Gruppo X	Gruppo Y
Media	5.00000	5.00000
Mediana	5.00000	2.00000
Varianza	6.66667	67.11111
Scarto quadratico medio	2.58199	8.19214
Coefficiente di variazione	0.51640	1.63842

Per quanto concerne la concentrazione, essendo i singoli dati già ordinati, per ottenere il rapporto di concentrazione R basta calcolare le seguenti quantità, per X e per Y, rispettivamente:

$$\sum_{i=1}^{9} x_i = 1 + 2 + \ldots + 9 = 45; \quad \sum_{i=1}^{9} i\,x_{(i)} = 1 \times 1 + 2 \times 2 + \ldots + 9 \times 9 = 285;$$

$$\sum_{i=1}^{9} y_i = 1 + 1 + \ldots + 28 = 45; \quad \sum_{i=1}^{9} i\,y_{(i)} = 1 \times 1 + 2 \times 1 + \ldots + 9 \times 28 = 347.$$

Alla fine, i due rapporti di concentrazione – utilizzando la formula semplificata, presentata prima dell'esempio 5.5 – risultano così determinati:

$$R_X = \frac{2}{8}\frac{285}{45} - \frac{10}{8} = \frac{1}{3} = 0.33333;$$

$$R_Y = \frac{2}{8}\frac{347}{45} - \frac{10}{8} = \frac{61}{90} = 0.67778.$$

Gli indici mostrano che il secondo gruppo presenta una concentrazione dei redditi molto più elevata del primo.

Allo scopo di fornire una rappresentazione geometrica di tale risultato, presentiamo il grafico della spezzata di concentrazione per i due gruppi (assieme alla retta di equidistribuzione) nella Figura 5.2. A tale scopo, i punti $(p_i, q_i), i = 1, 2, \ldots, 9$ sono stati ottenuti cumulando sequenzialmente per le ascisse le quote di unità e per le ordinate le corrispondenti quote di reddito posseduto.

La Figura 5.2 conferma che il secondo gruppo è molto più concentrato del primo perché l'area di concentrazione corrispondente è notevolmente più ampia. □

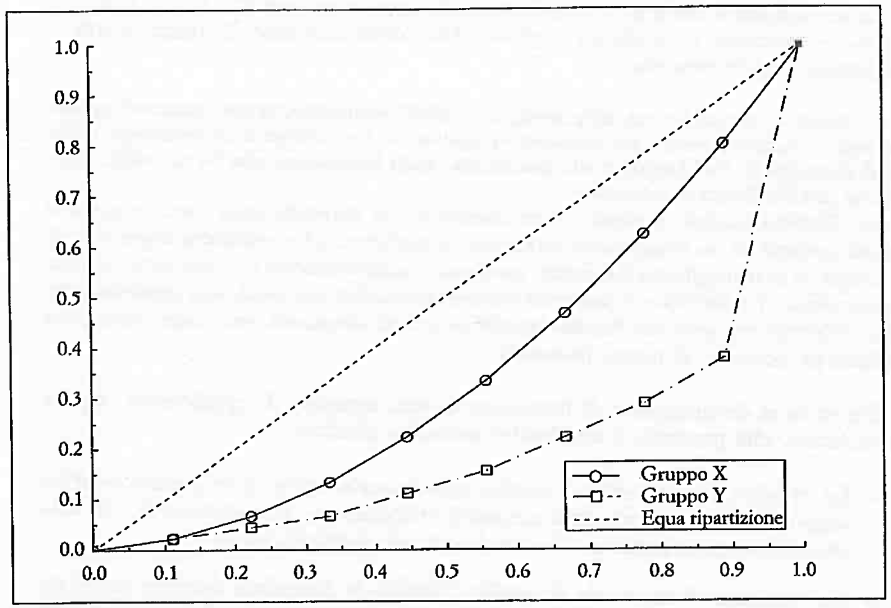

FIGURA 5.2. Curve di Lorenz per i dati dell'esempio 5.6.

Quando i dati sono disponibili in classi di modalità (come si verifica per le dichiarazioni dei redditi, per esempio) è necessario fare riferimento ad approssimazioni concernenti il calcolo dell'area di concentrazione. A tal proposito, una buona approssimazione di R (anche se per difetto) è fornita dalla formula seguente:

$$\tilde{R} = 1 - \sum_{i=0}^{n-1} (q_i + q_{i+1})(p_{i+1} - p_i) = \sum_{i=1}^{n-1} (p_i\, q_{i+1} - p_{i+1}\, q_i)\,.$$

Esempio 5.7

Supponiamo che i redditi dichiarati da due tipologie di professionisti siano quelli espressi nella tabella seguente, ove le classi di modalità sono espresse in migliaia di Euro annui. I valori e le classi sono qui riportati a puro scopo esemplificativo e non riflettono situazioni reali; tuttavia, la presentazione dei dati è quella usuale.

Classi di reddito	Gruppo A		Gruppo B	
	Numero di dichiaranti	Reddito dichiarato	Numero di dichiaranti	Reddito dichiarato
0 ⊢ – ⊣ 10	20	100	80	300
10 – ⊣ 20	30	500	30	450
20 – ⊣ 30	50	1200	20	500
30 – ⊣ 40	25	900	10	350
40 – ⊣ 50	15	800	5	225
50 – ⊣ 100	10	700	5	375
Totali	150	4200	150	2200

Mediante tali informazioni, deriviamo le quantità (p_i, q_i) che sono necessarie e sufficienti per il calcolo del rapporto di concentrazione secondo la formula approssimata per $\tilde{R}$. A tal fine, occorre costruire le tabelle cumulate del numero dei dichiaranti $\sum_j n_j$ e del reddito da loro posseduto $\sum_j x_{(j)}$, come evidenziato dalla tabella seguente, nella quale abbiamo inserito sia i valori assoluti cumulati che i valori relativi cumulati (in parentesi). Ovviamente, i valori relativi cumulati del numero dei dichiaranti sono le quantità p_i mentre i valori relativi cumulati del reddito sono le quantità q_i.

	Gruppo A				Gruppo B			
Classi di reddito	$\sum_j n_j$	(p_i)	$\sum_j x_{(j)}$	(q_i)	$\sum_j n_j$	(p_i)	$\sum_j x_{(j)}$	(q_i)
0 ⊢ – ⊣ 10	20	(0.133)	100	(0.024)	80	(0.534)	300	(0.136)
10 – ⊣ 20	50	(0.334)	600	(0.143)	110	(0.734)	750	(0.341)
20 – ⊣ 30	100	(0.667)	1800	(0.428)	130	(0.867)	1250	(0.569)
30 – ⊣ 40	125	(0.834)	2700	(0.643)	140	(0.934)	1600	(0.728)
40 – ⊣ 50	140	(0.934)	3500	(0.833)	145	(0.967)	1825	(0.828)
50 – ⊣ 100	150	(1.000)	4200	(1.000)	150	(1.000)	2200	(1.000)

Poiché nel nostro caso è $n = 5$, l'ultima espressione di $\tilde{R}$ diventa:

$$\tilde{R} = (p_1\, q_2 + p_2\, q_3 + p_3\, q_4 + p_4\, q_5) - (p_2\, q_1 + p_3\, q_2 + p_4\, q_3 + p_5\, q_4).$$

Applicandola ai due gruppi si ha, per il gruppo A:

$$\begin{aligned} \tilde{R} &= (0.133 \times 0.143 + 0.334 \times 0.428 + \ldots + 0.934 \times 1.000) + \\ &- (0.334 \times 0.024 + 0.667 \times 0.143 + \ldots + 1.000 \times 0.833) = 0.15886; \end{aligned}$$

mentre, per il gruppo B, si ha:

$$\begin{aligned} \tilde{R} &= (0.543 \times 0.341 + 0.734 \times 0.569 + \ldots + 0.967 \times 1.000) + \\ &- (0.734 \times 0.136 + 0.867 \times 0.341 + \ldots + 1.000 \times 0.828) = 0.51544\,. \end{aligned}$$

Come era presumibile, il gruppo B risulta sensibilmente più concentrato rispetto al gruppo A, perché – come era evidente sin dalla prima tabella – una frazione rilevante del reddito complessivo è posseduta da un numero ridotto di unità statistiche, quelle più ricche, collocate nelle ultime righe della tabella.

□

o Va ricordato che la concentrazione misura un aspetto della variabilità differente rispetto a quello misurato dalla varianza e dal coefficiente di variazione. Per esempio, se la variabile X si modifica in $X - c$, ove c è una qualsiasi costante, la concentrazione

misurata da R diminuisce se $c > 0$, così come il coefficiente di variazione, mentre la varianza di X non cambia. Ancora, se la variabile X si modifica in cX, allora il rapporto di concentrazione R ed il coefficiente di variazione CV non cambiano, ma la varianza si moltiplica per c^2.

o La scuola statistica italiana ha proposto innumerevoli altre misure di concentrazione per misurare, visualizzare, confrontare tale concetto, il quale è rilevante nelle situazioni di disuguaglianza o benessere socio-economico. Tali discussioni sono connesse a numerose altre, sia di natura metodologica che di natura applicata. Nel caso del reddito, per esempio, non è possibile immaginare che tutta la popolazione possieda reddito nullo (è necessario ipotizzare un reddito minimo di sopravvivenza) oppure che uno solo possieda l'intero ammontare della ricchezza disponibile (esiste un limite alla possibilità di accumulo). Analoghi problemi si pongono per la concentrazione degli investimenti, della popolazione sul territorio, della durata delle telefonate, e così via.

o Tra le utilizzazioni del concetto e degli indici di concentrazione, occorre ricordare il dibattito sulla misura della povertà che è strettamente connesso all'analisi della diseguaglianza perché, storicamente, a situazioni di grande diseguaglianza economica si accompagnano sempre situazioni di estrema povertà della popolazione. Ci limitiamo a segnalare la cosiddetta *linea di povertà* mediante la quale *si definisce povera una famiglia di due componenti la cui spesa per componente è inferiore alla metà della spesa pro-capite della intera popolazione di riferimento.*

5.6 Indici di eterogeneità

La discussione sin qui sviluppata ha riguardato la variabilità di un fenomeno che si presenta con modalità quantitative, per cui non è possibile applicare tale approccio a variabili di tipo qualitativo. Inoltre, esiste un concetto di variabilità che riguarda il modo con cui le frequenze si dispongono rispetto a situazioni estreme.

Per rispondere a tali problematiche, Gini nel 1912 ha definito la *mutabilità* come «l'*attitudine di un carattere ad assumere differenti modalità qualitative*». Così, la mutabilità è la possibilità di variare per una variabile qualitativa tra una perfetta omogeneità (quando la variabile si manifesta mediante un solo attributo) e una qualche eterogeneità (se nella popolazione vi sono almeno due attributi differenti).

o Per esempio, nel data set STUDENTI, la variabile qualitativa «Diploma di provenienza» assume i quattro attributi: Classico (CL), Scientifico (SC), Tecnico-Professionale (TE), Altri (AL).

Allora, il fenomeno sarebbe *totalmente disomogeneo* se la popolazione si distribuisse equamente tra le quattro modalità qualitative (ogni Diploma sarebbe presente esattamente per il 25%). All'opposto, risulterebbe *completamente omogeneo* se tutta la popolazione considerata possedesse un solo tipo di Diploma.

Nei casi reali, evidentemente, vi sarà una maggiore o minore accentuazione di un attributo rispetto agli altri, per cui il confronto tra le distribuzioni avviene mediante misure che fanno riferimento agli estremi di massima o minima omogeneità ovvero, di minima o massima eterogeneità.

La *eterogeneità* misura la variabilità delle frequenze relative $f_i, i = 1, 2, \ldots, k$, senza coinvolgere le modalità o gli attributi della variabile X, quantitativa o qualitativa, rispettivamente.

o Anche se introdotta con riferimento a variabili qualitative, la discussione di questo paragrafo è possibile anche per caratteri quantitativi. La problematica coinvolge molti settori disciplinari, dall'Ecologia alla Sociologia, dalla Linguistica alla Demografia, dalla Biologia alla Statistica economica.

La diversità razziale presente in un continente, la diversità delle specie animali e vegetali presenti in un determinato territorio, la ricchezza del vocabolario linguistico di un Autore, la disuguaglianza dei redditi personali, l'accentramento di quasi tutti i voti in un solo partito, l'indecisione di fronte ad opzioni alternative, etc. sembrano problematiche molto differenti tra loro ma l'analisi statistica di tali situazioni può essere perseguita mediante gli indicatori di questo paragrafo.

Sia data la distribuzione di frequenza di una variabile X, qualitativa oppure quantitativa, che presenta k attributi o modalità distinte.

- La *minima eterogeneità* si verifica allorquando tutte le frequenze relative sono accentrate in una sola modalità (diciamo x_1, per semplicità): in tale situazione, ci attendiamo che l'indice di eterogeneità valga 0.
- La *massima eterogeneità* si verifica quando le frequenze relative si distribuiscono in parti esattamente eguali tra le k modalità (cioè la frequenza relativa di ciascun attributo o modalità di X è sempre uguale a $\frac{1}{k}$): in tale situazione, ci attendiamo che l'indice di eterogeneità raggiunga il massimo.

Poiché le misure di eterogeneità sono connesse a problematiche molto diverse tra loro, è comune trovare lo stesso indice con denominazioni differenti ed attribuito a differenti Autori. Tuttavia, gli indici di eterogeneità più diffusi sono quelli di Gini (1912) e di Shannon (1948). Nella versione assoluta, ciascuno di essi varia tra 0 ed il proprio massimo (indicato in tabella); in quella normalizzata, tali indici variano sempre tra 0 e 1.

Autore	Indice assoluto	Valore massimo	Indice normalizzato
Gini	$G = 1 - \sum_{i=1}^{k} f_i^2$	$G_{max} = 1 - \frac{1}{k}$	$G^* = \frac{k\,G}{k-1}$
Shannon	$H = -\sum_{i=1}^{k} f_i \log(f_i)$	$H_{max} = \log(k)$	$H^* = \frac{H}{\log(k)}$

o Sfruttando lo sviluppo in serie della funzione logaritmica, si dimostra che – se le frequenze relative sono molto piccole – allora: $H \simeq G$, per cui al di là delle inevitabili approssimazioni numeriche, questi due indici sono espressioni concordi della eterogeneità.

o Per migliorarne la capacità selettiva, Leti e Frosini hanno proposto altre misure che sono comunque trasformazioni degli indici precedenti e di cui forniamo nel riquadro seguente le sole versioni normalizzate.

Autore	Indice normalizzato
Leti	$L^* = \frac{exp(H)}{k-1}$
Frosini	$F^* = 1 - \sqrt{1 - G^*}$

o Tutte le misure di eterogeneità qui presentate hanno il comune difetto di variare in ambiti molto ristretti per cui, spesso, sono poco efficaci per discriminare tra situazioni reali quando queste non siano già molto diverse fra loro.

Esempio 5.8

Confrontiamo i risultati di due consultazioni elettorali, avvenute *prima* e *dopo* una legge di riforma elettorale che puntava a ridurre la polverizzazione tra le formazioni partitiche, al fine di assicurare maggiore governabilità al Paese. Supponiamo, quindi, che la distribuzione dei voti sia sintetizzata nella tabella seguente, dove sono riportate in ordine alfabetico le proporzioni di preferenze elettorali (cioè le frequenze relative) attribuite a ciascun partito politico.

Partiti politici	Frequenze relative (f_i)	
	Prima della legge	Dopo la legge
A	0.02	0.02
B	0.08	0.10
C	0.30	0.20
D	0.40	0.35
E	0.04	0.10
F	0.10	0.15
G	0.06	0.08
Totali	1.00	1.00

Mediante tali informazioni, il calcolo dell'indice di eterogeneità di Gini è immediato:

$$\begin{aligned} G_{prima} &= 1 - (0.02^2 + 0.08^2 + \ldots + 0.06^2) = 0.7280; \\ G_{dopo} &= 1 - (0.02^2 + 0.10^2 + \ldots + 0.08^2) = 0.7882. \end{aligned}$$

Analogamente, per il calcolo dell'indice di Shannon si ottiene:

$$\begin{aligned} H_{prima} &= -0.02 \times \log(0.02) - 0.08 \times \log(0.08) - \ldots - 0.06 \times \log(0.06) = 1.53583; \\ H_{dopo} &= -0.02 \times \log(0.02) - 0.10 \times \log(0.10) - \ldots - 0.08 \times \log(0.08) = 1.71471. \end{aligned}$$

Da tali indici, essendo il numero di attributi pari a $k = 7$, si deducono le corrispondenti misure normalizzate, che riassumiamo nella tabella seguente, insieme agli indici di Leti e Frosini:

Indici normalizzati	Prima della legge	Dopo la legge
Gini (G^*)	0.84933	0.91957
Shannon (H^*)	0.78926	0.88119
Leti (L^*)	0.20029	0.23563
Frosini (F^*)	0.61184	0.71640

Tutte le misure utilizzate sono concordi nell'evidenziare che dopo la legge vi è stata una maggiore eterogeneità, e quindi una più accentuata polverizzazione delle scelte elettorali. In questo esempio, i dati mostrano un effetto contrario a quello auspicato dalla riforma elettorale.

□

5.7 Considerazioni finali

La variabilità dei fenomeni del mondo reale gioca un ruolo centrale per l'intera analisi statistica, da molteplici punti di vista.

In primo luogo, la variabilità è lo stimolo primario che genera e muove l'indagine scientifica, perché la diversità pone domande e cerca risposte.

o Ciò avviene nella vita quotidiana e nella pianificazione aziendale, nel miglioramento dei processi produttivi e nella previsione di dati economici. In fin dei conti, è la presenza di un mondo mutevole e diversificato che sollecita problemi sulle origini della variabilità, sulla connessione tra fenomeni e sulle strategie ottimali per affrontare tali variazioni. Se il mondo che ci circonda e le situazioni che quotidianamente si incontrano non fossero caratterizzate da modalità ed attributi variabili, l'essere umano non avrebbe necessità di prendere decisioni e, alla fine, di imparare, evolversi, migliorare il suo status, fare progetti e realizzarli. Per questo, da un punto di vista epistemologico, *la variabilità è strettamente connessa alla INFORMAZIONE* perché grazie ad essa possiamo verificare ipotesi, classificare e discriminare le osservazioni, prendere decisioni, graduare efficacemente le scelte.

In secondo luogo, la variabilità presente sia nei risultati sperimentali che nella vita quotidiana genera insicurezza, paure, rischi e, quindi, errori nel rapportarsi dell'Uomo rispetto al tempo, al territorio, a circostanze differenti.

o Per questo, *la variabilità è anche INCERTEZZA* che produce la tendenza molto naturale ad affidarsi a fenomeni poco mutevoli o, almeno, che possiedano una dispersione controllabile. I meccanismi di tutela, difesa, sicurezza ed assicurazione sono sostanzialmente delle strategie rivolte al contenimento della variabilità dei fenomeni del mondo reale.

Come vedremo, il punto di riferimento più idoneo per lo studio della variabilità è quello dei modelli probabilistici nel cui ambito sia l'informazione che l'incertezza assumono una interpretazione ed un uso efficienti, secondo paradigmi formali, rigorosi ed utilizzabili a fini decisionali.

Ambiente R

Indici statistici di variabilità

R5.1 Variabilità rispetto ad un centro

Occorre considerare che (in aderenza ad alcuni risultati inferenziali che saranno discussi nel paragrafo 12.5 del testo), R utilizza $(n-1)$ al denominatore della varianza, mentre la formula corretta quando trattasi di una popolazione prevede n al denominatore. Per questo, la funzione `var()` calcolata su una serie di n modalità va moltiplicata per $(n-1)/n = 1 - 1/n$ per ottenere la ***varianza*** secondo la formula definita nel Capitolo 5 del testo. In alternativa, si può definire (e poi utilizzare) la seguente funzione `varpop()`:

```
R==> varpop<-function(variabile){
          var(variabile)*(1-1/length(variabile))
}
```

Per esempio, la varianza della variabile `lunghezza` del data set `camellia` si ottiene con le istruzioni:

```
R==> dati=read.table("camellia.txt", header=TRUE);
R==> lunghezza=dati$lunghezza;

R==> var(lunghezza); varpop(lunghezza);
[1] 171.11
[1] 170.59
```

oppure, in alternativa, mediante:

```
R==> with(dati, var(lunghezza));
[1] 171.11
R==> with(dati,varpop(lunghezza));
[1] 170.59
```

Come si vede, la differenza tra le due quantità non è eccessiva (in effetti, tende a 0 al crescere di n) ma, in piccole popolazioni, può essere sensibile.

Gli indici ***scarto quadratico medio*** σ e ***coefficiente di variazione*** CV si ottengono mediante:

```
R==> sqmlun=sqrt(varpop(lunghezza)); cvlun=sqm/mean(lunghezza);
R==> sqmlun; cvlun;
[1] 13.061
```

```
[1] 0.13387
R==> sqmlar=sqrt(varpop(larghezza)); cvlar=sqm/mean(larghezza);
R==> sqmlar; cvlar;
[1] 6.7964
[1] 0.28569
```

Nel data set considerato risulta che lo scarto quadratico medio della variabile lunghezza è circa il 13% della sua media mentre per larghezza è quasi il 30% della sua media. In questo esempio, l'indice assoluto (=*sqm*) avrebbe fornito la falsa impressione di maggiore variabilità di lunghezza mentre quello relativo (=*CV*) mostra che, rispetto al livello medio, larghezza varia di più. Ciò conferma la caratterizzazione della forma della foglia del data set camellia che, appunto, consta di due specie che si differenziano più per la larghezza che non per la lunghezza.

Utilizzando la funzione abs() (che fornisce il valore assoluto del suo argomento), lo *scostamento semplice dalla media* ssmedia e lo *scostamento semplice dalla mediana* ssmediana si calcolano nel modo seguente;

```
R==> ssmedia=mean(abs(lunghezza-mean(lunghezza)));
R==> ssmediana=mean(abs(lunghezza-median(lunghezza)));
R==> ssmediana; ssmedia; sqmlun;
[1] 10.312
[1] 10.334
[1] 13.061
```

Si osservi come è confermato il previsto ordinamento, citato nel testo, fra questi tre assoluti indici di variabilità.

R5.2 Variabilità dalle modalità ordinate

Per il *campo di variazione* cvar (già introdotto), il *campo di variazione interdecile* idr ed il *campo di variazione interquartile* iqr si possono utilizzare le seguenti definizioni, che applichiamo alla variabile larghezza nel data set camellia:

```
R==> dati=read.table("camellia.txt", header=TRUE);
R==> larghezza=dati[,3];
R==> cvar=max(larghezza)-min(larghezza); cvar;
[1] 36

R==> dec=quantile(larghezza,probs=(1:9)/10);
R==> idr=dec[9]-dec[1]; idr;
90%
 16

R==> qua=quantile(larghezza,probs=c(0.25,0.50,0.75));
R==> iqr=qua[3]-qua[1]; iqr;
75%
  9
```

Si ricordi che i tre indici misurano la differenza tra le modalità ordinate all'interno delle quali sono collocati, rispettivamente, il 100%, l'80% e il 50% di tutte le modalità (per cui le percentuali che R indica in output si riferiscono alla percentuale dell'ultimo quantile utilizzato).

Infine, si noti che R possiede una propria funzione IQR() per calcolare il *campo di variazione interquartile*.


```
R==> IQR(lunghezza); IQR(larghezza);
[1] 17
[1] 9
```

R5.3 Mutua variabilità

Per il calcolo della ***differenza semplice media*** Δ applichiamo la formula del testo che utilizza le modalità ordinate proponendo una funzione Delta appositamente costruita.

```
R==> Delta<-function(variabile){
          n=length(variabile); media=mean(variabile);
          varsort=sort(variabile);
          (4*sum((1:n)*varsort)/n-2*media*(n+1))/(n-1)
}
```

È quindi immediato il calcolo del ***rapporto di concentrazione di Gini*** secondo la formula: $R = \Delta/(2\mu)$

```
R==> gini=Delta(variabile)/(2*mean(variabile))
```

L'esempio ne mostra l'applicazione per la variabile componenti nel data set studenti:

```
### Legge il data set STUDENTI
R==> dati=read.table("studenti.txt", header=TRUE);

R==> with(dati, mean(componenti));
[1] 4.5125       ### media componenti

R==> with(dati, Delta(componenti));
[1] 1.1867       ### differenza semplice media

R==> componenti=dati[,10];
R==> gini=Delta(componenti)/(2*mean(componenti)); gini;
[1] 0.13149      ### rapporto concentrazioni di Gini
```

Quindi, il numero medio di componenti per famiglia è 4.51 con una differenza media (reciproca, tra le famiglie cioè) di circa 1.2 componenti; la concentrazione è molto bassa (il 13% di quella massima ipotizzabile). In tale esempio, la situazione di massima concentrazione è puramente astratta perché si raggiungerebbe nell'ipotesi di una sola famiglia composta da tutti i componenti del data set (pari a 12450 componenti!).

R5.4 Curva di Lorenz

La curva di Lorenz visualizza il rapporto di concentrazione di Gini mediante una rappresentazione grafica che pone in ascissa la funzione di ripartizione empirica ed in ordinata la quota (rispetto al totale) dei valori cumulati della variabile trasferibile. Il *rapporto di concentrazione di Gini* si ottiene (graficamente) come rapporto tra l'area sottesa tra la retta di equidistribuzione e la curva di Lorenz e l'area massima definita dal triangolo inferiore del quadrato unitario.

Implementiamo una funzione che visualizza sia il valore di R che il grafico della curva di Lorenz. Dall'esame delle istruzioni utilizzate, il lettore potrà dedurre agevolmente alcuni strumenti grafici di R (per esempio, come collocare in un testo il risultato di un calcolo

come l'indice R oppure come apporre una scritta orientata a 43°, in questo caso, e così via).

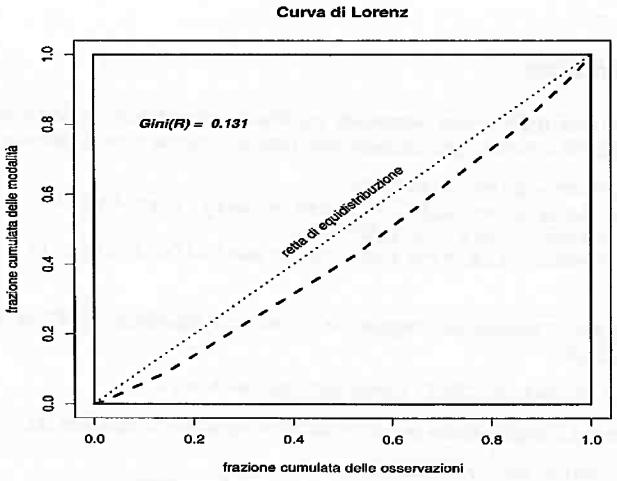

FIGURA R5.1. *Curva di Lorenz per la variabile componenti del data set studenti.*

```
Lorenz<-function(x){
        n=length(x); x=sort(x);
        gini=0.5*Delta(x)/mean(x);
        ascisse=(0:n)/n; media=mean(variabile);
        ordinate=c(0,cumsum(x)/sum(x));
        plot(ascisse,ascisse,type="n",xlim=c(0,1),ylim=c(0,1),
             xlab="frazione cumulata delle osservazioni",
             ylab="frazione cumulata delle modalità");
        title(main="Curva di Lorenz");
        lines(ascisse,ascisse,lty=3,lwd=2);
        lines(ascisse,ordinate,lty=2,lwd=3);
        segments(0,0,1,0,lwd=2);segments(1,0,1,1,lwd=2);
        segments(1,1,0,1,lwd=2);segments(0,1,0,0,lwd=2);
        text(0.2,0.8,paste("Gini(R) = ", round(gini,digit=3)),
            font=4);
        text(0.5,0.55,"retta di equidistribuzione",font=2,srt=43);
}
```

Calcoliamo la concentrazione della variabile componenti nel data set studenti, mediante le istruzioni:

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> with(dati, Lorenz(componenti));
```

ottenendo la Figura R 5.1. Si conferma la bassa concentrazione di tale variabile evidenziata da un'area di concentrazione piuttosto ridotta.

R5.5 Indici di eterogeneità

L'indice di eterogeneità del Gini e di Shannon sono definiti per qualsiasi variabile (qualitativa o quantitativa) tramite la sua distribuzione di frequenza. Nel caso di una variabile discreta (come la variabile componenti che assume modalità distinte da 1 a $m = 12$), l'*indice di eterogeneità di Gini* e quello *di Frosini* si ottengono nel modo seguente:

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> componenti=dati[,10]; emme=max(componenti);
R==> freq=tabulate(componenti,nbins=emme);
R==> freq=freq/sum(freq);                    # frequenze relative

### indice normalizzato di Gini
R==> hetergini=(1-sum(freq^2))*m/(m-1);
### indice normalizzato di Frosini
R==> frosini=(1-sqrt(1-hetergini));
R==> hetergini; frosini;
[1] 0.79054
[1] 0.54234
```

Come si vede, invece di calcolarlo a partire dalle frequenze, l'indice di Frosini è ottenuto da quello di Gini (mediante la relazione riportata nel testo).

Per il calcolo dell'*entropia di Shannon* (e *di Leti*), invece, bisogna evitare che nella somma intervengano termini con frequenza nulla in quanto il termine $0*log(0)$ è indefinito mentre R attribuisce il valore 0 all'espressione $\log(0^0)$ (in effetti, è: $\lim_{x\to 0+} \log(x^x) = 0$). Per questo, anche se computazionalmente è inefficiente (perché l'espressione $-\sum_i \log\left(f_i^{f_i}\right)$ coinvolge numeri molto elevati), utilizzeremo l'espressione che evita in R il rischio di quantità indeterminate.

Nell'esempio seguente le variabili sono quelle precedentemente definite:

```
R==> freq=freq/sum(freq);                    # frequenze relative
R==> shannon=-sum(log(freq^freq));           # indice di Shannon
R==> shannon=shannon/log(emme);              # indice normalizzato
                                             # di Shannon
R==> leti=exp(shannon)/(emme-1);             # indice normalizzato di Leti
R==> shannon; leti;
[1] 0.6201
[1] 0.1690
```

Capitolo sesto

Forma di una distribuzione di frequenza

6.1 Introduzione

La posizione e la variabilità di una distribuzione di frequenza non esauriscono le informazioni contenute nei dati, poiché due variabili statistiche possono avere la stessa posizione e la stessa variabilità ma differire per il peso dei valori più grandi o più piccoli rispetto al valore centrale, a causa del comportamento differenziato delle «code» della distribuzione.

Questi aspetti possono essere riferiti a formulazioni teoriche oppure si può limitare l'analisi alla descrizione della forma dell'istogramma, cercando anche qualche visualizzazione alternativa ma sintetica della distribuzione di frequenza.

Nel seguito discuteremo alcune misure concernenti la forma di una distribuzione (tradizionalmente, chiamate di asimmetria e di curtosi) e introdurremo poi alcune rappresentazioni dell'istogramma che consentono in modo semplice ed immediato la sintesi ed il confronto delle informazioni disponibili sulle variabili statistiche.

6.2 Asimmetria di una distribuzione di frequenza

Un primo aspetto della forma di una distribuzione riguarda l'asimmetria più o meno accentuata in rapporto ad un punto centrale.

Se la media supera la mediana, significa che la distribuzione si «attarda» verso valori alti e quindi superiori alla mediana, per cui siamo in presenza di una distribuzione che presenta in proporzione più valori (si dice «una coda») verso il semiasse positivo delle x, cioè verso destra: si parla, allora, di *asimmetria positiva* (a destra).

Per contro, se la media è inferiore alla mediana, saremo in presenza di una distribuzione di frequenza che presenta una «coda» verso sinistra e, quindi, si parla di *asimmetria negativa* (a sinistra).

Supponendo che la distribuzione di una variabile statistica abbia una sola moda, è agevole convincersi che:

- ***Asimmetria positiva*** $\Longrightarrow Moda < Mediana < Media$
- ***Asimmetria negativa*** $\Longrightarrow Media < Mediana < Moda$

Un primo indice di asimmetria può essere ottenuto mediante la differenza tra media e mediana. Poiché è stato dimostrato che tale differenza non può superare, in valore assoluto, lo scarto quadratico medio della variabile, un ***indice normalizzato di asimmetria*** che assume valori nell'intervallo $[-1, 1]$ è definito da:

$$\mathcal{A} = \frac{\mu - Me}{\sigma} .$$

Una distribuzione con asimmetria positiva presenterà un indice $\mathcal{A} > 0$; una distribuzione con asimmetria negativa presenterà un indice $\mathcal{A} < 0$; una distribuzione simmetrica presenterà un indice $\mathcal{A} = 0$.

Un approccio più generale sostiene che qualsiasi indice di forma della distribuzione dovrebbe prescindere dalla posizione e dalla variabilità per consentire confronti tra fenomeni di natura diversa. Il che è possibile tramite la definizione di variabile standardizzata, come ora vedremo.

Per ogni variabile X, definiamo la ***variabile standardizzata*** $Z = Z(X)$ mediante la seguente trasformazione lineare (cioè di primo grado in X):

$$Z = \frac{X - \mu}{\sigma} .$$

- Anche a prescindere da ogni discussione sulla forma, l'introduzione della variabile standardizzata ha molte applicazioni in Statistica e di per sé consente il confronto di più distribuzioni con medie e varianze differenti. In particolare, l'uso dei dati standardizzati può aiutare nella determinazione di valori anomali ovvero nella specificazione di estremi della distribuzione. Qui di seguito, ci limitiamo a discuterne l'uso soprattutto con riferimento alla proposta di indici di forma di una distribuzione di frequenza per una variabile quantitativa.
- È immediato convincersi che – trattandosi di una traslazione rispetto ad una costante e di una successiva divisione per un'altra costante – la variabile Z ha una distribuzione di frequenza che presenta la stessa forma di quella originaria di X, dalla quale si differenzia, appunto, solo per la posizione e la scala.
- D'altra parte, la variabile Z *non dipende dall'unità di misura* di X ed inoltre si dimostra che essa ha media aritmetica sempre uguale a 0 e varianza sempre uguale a 1.

Infatti, la serie $(x_1, x_2, \ldots, x_n)$ delle modalità per la variabile X mediante la trasformazione $z_i = \dfrac{x_i - \mu}{\sigma}, i = 1, 2, \ldots, n$ diventa:

$$(x_1, x_2, \ldots, x_n) \Longrightarrow \left(\frac{x_1 - \mu}{\sigma}, \frac{x_2 - \mu}{\sigma}, \ldots, \frac{x_n - \mu}{\sigma}\right) = (z_1, z_2, \ldots, z_n) .$$

Allora, media e varianza di Z sono derivabili dai seguenti passaggi:

$$\mu_Z = \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right) = \frac{1}{\sigma}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)\right] = \frac{1}{\sigma}0 = 0 .$$

$$\sigma_Z^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i - \mu_Z)^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i - 0)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2 =$$

$$= \frac{1}{\sigma^2}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2\right] = \frac{1}{\sigma^2}\sigma^2 = 1 .$$

Per questo, essendo dotata di misure *standard* di posizione e variabilità, la variabile standardizzata Z può essere il punto di partenza per costruire idonee misure della forma di una distribuzione di frequenza. Infatti, due distribuzioni di frequenza standardizzate avranno la stessa media (0) e la stessa varianza (1), per cui possono essere «sovrapposte» graficamente per una loro valutazione comparata.

Su tali basi, l'***indice di asimmetria di Fisher*** è definito come la media aritmetica delle terze potenze della variabile standardizzata:

$$\gamma = \frac{1}{n}\sum_{i=1}^{n}(z_i)^3 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^3 = \frac{1}{n}\sum_{i=1}^{k}\left(\frac{x_i - \mu}{\sigma}\right)^3 n_i ,$$

dove l'ultima espressione dell'indice si applica quando i dati sono disponibili in seriazione.

L'indice γ è positivo, negativo o nullo per una distribuzione, rispettivamente, asimmetrica positiva, negativa o perfettamente simmetrica.

Esempio 6.1

Supponiamo di avere due variabili X e Y le cui osservazioni sono raccolte ed organizzate nelle prime due colonne (modalità e frequenze) delle tabelle seguenti e graficamente rappresentate nei diagrammi a barre della Figura 6.1.

Per tali distribuzioni, si calcolano media, varianza, e scarto quadratico medio (s.q.m.) che risultano essere:

Variabili	Media (μ)	Varianza (σ^2)	s.q.m. (σ)
X	7.30	9.41000	3.06758
Y	5.18	20.98760	4.58122

Si tratta di calcolare, per ciascuna di esse, le rispettive variabili standardizzate, sottraendo dalle modalità le medie e poi dividendo per lo scarto quadratico medio. Le modalità standardizzate elevate alla terza potenza sono indicate nella terza colonna delle due tabelle e, moltiplicate per le frequenze assolute, forniscono le quantità necessarie per calcolare il coefficiente di asimmetria γ.

Variabile X		
x_i	n_i	$\left(\frac{x_i - 7.30}{3.06758}\right)^3 n_i$
1	10	$(-2.05374)^3 \times 10$
2	20	$(-1.72775)^3 \times 20$
3	30	$(-1.40176)^3 \times 30$
5	40	$(-0.74978)^3 \times 40$
7	50	$(-0.09790)^3 \times 50$
10	150	$(0.88018)^3 \times 150$
Totali	300	-187.02980

Variabile Y		
y_i	n_i	$\left(\frac{y_i - 5.18}{4.58122}\right)^3 n_i$
2	160	$(-0.69414)^3 \times 160$
3	110	$(-0.47586)^3 \times 110$
4	90	$(-0.25758)^3 \times 90$
8	60	$(0.61556)^3 \times 60$
10	50	$(1.05212)^3 \times 50$
20	30	$(3.23495)^3 \times 30$
Totali	500	$+1020.92250$

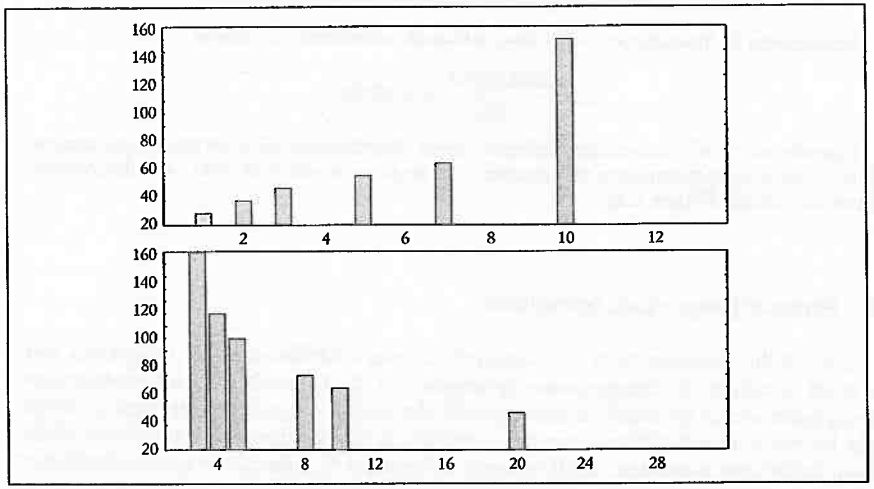

FIGURA 6.1. Diagramma a barre per le variabili dell'esempio 6.1.

Dividendo le somme delle modalità standardizzate alla terza potenza per le rispettive numerosità si ottengono gli indici di asimmetria, che sono:

$$\gamma_X = -0.62343; \qquad \gamma_Y = +2.04184\ .$$

Essi esprimono che la distribuzione di X ha una coda a sinistra (è asimmetrica negativa), mentre la distribuzione di Y ha una coda a destra (è asimmetrica positiva). □

Va ribadita, qui, una cautela di carattere generale che ha accompagnato l'intera discussione sui rischi derivanti dalla sintesi di una distribuzione di frequenza mediante un solo valore. Infatti, se ad un comportamento ben definito corrisponde un indice di un certo valore o di un certo segno, quell'indice non deriva necessariamente da distribuzioni univocamente determinate dal valore o dal segno dell'indice stesso.

L'esempio seguente conferma, in particolare, la necessità di un atteggiamento critico verso tutti gli indici statistici e, in particolare, verso gli indici di asimmetria.

Esempio 6.2

Siano X e Y due variabili le cui distribuzioni di frequenza sono presentate nelle tabelle seguenti.

Variabile X		Variabile Y	
x_i	n_i	y_i	n_i
1	7	1	1
2	6	2	2
3	5	3	3
4	4	4	4
5	3	5	5
6	2	6	6
7	1	7	7

Le due distribuzioni sono l'una speculare dell'altra e, precisamente, X presenta una netta asimmetria positiva mentre Y evidenzia una altrettanto netta asimmetria negativa. Ciò è confermato dai valori dei corrispondenti indici γ che risultano essere pari a:

$$\gamma_X = +\frac{1}{\sqrt{3}} = +0.57735\,; \quad \gamma_Y = -\frac{1}{\sqrt{3}} = -0.57735\,.$$

Tuttavia, se si fosse utilizzato l'indice $\mathcal{A}$ per valutare l'asimmetria delle due distribuzioni si sarebbe giunti ad una palese contraddizione. Infatti, la media di X coincide con la mediana di X, e lo stesso vale per Y:

Variabili	Media	Mediana	Indice $\mathcal{A}$
X	3	3	0
Y	5	5	0

Come si vede, l'indice $\mathcal{A}$ vale 0 per entrambe le variabili, anche se le corrispondenti distribuzioni *non sono simmetriche.* □

6.3 Curtosi di una distribuzione di frequenza

Un altro aspetto della forma di una distribuzione di frequenza riguarda il maggiore o minore appuntimento e, conseguentemente, il peso più o meno accentuato delle code rispetto alla parte centrale della distribuzione.

o Per motivi che saranno evidenti nella parte seconda, molti fenomeni reali tendono a distribuirsi in modo da assumere valori più frequenti in corrispondenza dei valori centrali della distribuzione. È allora importante misurare il rapporto relativo tra «corpo centrale» e «code» di una distribuzione di frequenza, e tale rapporto è connesso all'appiattimento o

a picchi accentuati nella distribuzione di frequenza. Questo aspetto di una distribuzione, con termine di origine greca, si definisce *curtosi*.

o Effettuando la media delle potenze quarte della variabile standardizzata Z, si ottiene un indice statistico che deve la sua fortuna alla circostanza per cui esso vale esattamente 3 per una importante distribuzione teorica (quella Gaussiana, di cui ci occuperemo nel paragrafo 10.3).

Su tali basi, l'***indice di curtosi di Pearson*** è definito come la media aritmetica delle potenze quarte della variabile standardizzata:

$$\beta = \frac{1}{n}\sum_{i=1}^{n}(z_i)^4 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^4 = \frac{1}{n}\sum_{i=1}^{k}\left(\frac{x_i-\mu}{\sigma}\right)^4 n_i\,,$$

dove l'ultima espressione dell'indice si applica quando i dati sono disponibili in seriazione.

Per tale indice sono valide alcune proprietà, tra le quali elenchiamo le seguenti.

1. Si dimostra che: $\beta \geq 1$.

2. Per una particolare distribuzione teorica (quella Gaussiana, di cui al paragrafo 10.3) si dimostra che $\beta = 3$.

3. Per distribuzioni di frequenza relativamente più «appuntite» l'indice β è maggiore di 3 (tale distribuzione è detta *leptocurtica*).

4. Per distribuzioni tendenzialmente più «piatte» l'indice β è inferiore a 3 (tale distribuzione è detta *platicurtica*).

o Talvolta, si utilizza l'*indice di curtosi di Fisher* che vale $\beta - 3$.

o Gli indici di curtosi vanno interpretati con particolare cautela nei casi in cui le distribuzioni sono fortemente asimmetriche ovvero sono pluri-modali.

Esempio 6.3

Presentiamo nella tabella seguente i calcoli occorrenti per determinare l'indice di curtosi β per la variabile X = «Misura delle scarpe» di 1230 studentesse, già discussa nell'esempio 4.2. Per tale distribuzione si ha: $\mu = 37.81870$; $\sigma = 1.30744$.

x_i	n_i	$\frac{x_i - 37.81870}{1.30744}$	$\left(\frac{x_i - 37.81870}{1.30744}\right)^4 n_i$
34	2	−2.92074	145.54628
35	35	−2.15589	756.08791
36	148	−1.39104	554.13367
37	327	−0.62618	50.27528
38	359	0.13867	0.13274
39	227	0.90352	151.27850
40	115	1.66837	890.98299
41	13	2.43322	455.69223
42	3	3.19808	313.81683
43	1	3.96293	246.64059
Totali	1230	------	3564.58703

Applicando la formula per i dati disponibile in seriazione, si ottiene:

$$\beta = \frac{3564.58703}{1230} = 2.89804\,.$$

Essendo vicino a 3, tale indice conferma che la distribuzione di X ha una forma campanulare ed approssimativamente simmetrica attorno ai valori centrali (come dal diagramma a barre di cui alla Figura 4.1).

□

6.4 Forma e istogramma perequato

Una delle rappresentazioni standard di una distribuzione di frequenza per variabili continue è l'istogramma (paragrafo 3.2); tuttavia, la sua costruzione presuppone alcuni elementi di arbitrarietà che possono condizionare ogni giudizio sulla forma di una distribuzione: per esempio, la determinazione del numero delle classi, della loro ampiezza, degli estremi del campo di definizione (che potrebbero anche non coincidere con i minimi ed i massimi osservati).

o Per la determinazione del numero di classi, fra le tante proposte, una soluzione resistente è quella di Friedman e Diaconis che suggeriscono di determinare il numero k delle classi di un istogramma mediante la relazione:

$$k = \frac{max(X) - min(X)}{2\,IQR}\sqrt[3]{n}\,,$$

dove IQR è il campo di variazione interquartile di X.

Un approccio preferibile rispetto all'istogramma classico, e che riteniamo auspicabile si diffonda in futuro grazie alla disponibilità di software statistico di buona qualità, è quello dell'***istogramma perequato***.

o Ci limitiamo ad una presentazione informale ed intuitiva di tali idee. In breve, anziché fissare un numero di classi come nell'istogramma tradizionale – costruito sulle frequenze relative in modo che l'area complessiva racchiusa dall'istogramma sia pari ad 1 – l'istogramma perequato si ottiene sostituendo alla sequenza dei valori osservati $(x_1, x_2, \ldots, x_n)$ una funzione continua dell'ascissa ottenuta come media ponderata di un opportuno intorno dei valori $(x_1, x_2, \ldots, x_n)$.

La particolare funzione di ponderazione viene detta *nucleo* (*kernel*, in inglese) e la letteratura ha sviluppato numerose soluzioni a tale riguardo.

Si ottiene, così, una funzione regolare che consente una visualizzazione continua dell'andamento complessivo della distribuzione di frequenza e, soprattutto, della sua forma. Tale funzione migliora nettamente la qualità della rappresentazione delle osservazioni e consente confronti più efficaci.

Esempio 6.4

Dal data set STUDENTI confrontiamo i due istogrammi perequati della variabile «Peso» per Femmine (a sinistra) e Maschi (a destra) – che vengono presentati sovrapposti nella Figura R 6.1. Le istruzioni per ottenere i vari istogrammi sono riportate nel paragrafo R 6.2 assieme a tale figura.

Essi mostrano chiaramente che la distribuzione del Peso delle studentesse è non solo più spostato a sinistra ma anche meno variabile, a differenza di quello degli studenti. Si noti che, in tali situazioni, la sovrapposizione degli analoghi istogrammi distintamente per i due sessi – possibile in linea di principio – non avrebbe offerto una analoga chiara evidenza.

□

6.5 Metodi esplorativi e distribuzioni di frequenza

A partire dagli anni '60, sotto la spinta determinante di Tukey e degli statistici americani della sua scuola, si sono introdotte e diffuse tecniche statistiche di tipo descrittivo le quali, grazie alle recenti potenzialità di software grafico, consentono di rappresentare, persino in modo dinamico, il rapporto tra modalità ed indici statistici.

Tali studi – denominati EDA (=*Explorative Data Analysis*) – hanno avuto il merito, anche filosofico, di accentuare l'aspetto esplorativo della Statistica come prioritario rispetto a quello confermativo (tipico dell'inferenza e della modellistica), al fine di esplicitare dagli stessi dati le ipotesi utili e le correzioni da apportare.

Oggi, l'analisi esplorativa dei dati costituisce una branca della Statistica di ampio respiro e piuttosto impegnativa, con numerose proposte. Noi ci limitiamo a segnalare (nella formulazione essenziale) solo quella che costituisce, a nostro parere, una visualizzazione sintetica e resistente, ma anche semplice ed efficace della distribuzione di frequenza per una variabile quantitativa.

Tukey nel 1977 introdusse una rappresentazione denominata «grafico a scatola», universalmente chiamata ***box-plot*** (come anche noi diremo nel seguito) ed un esempio del quale, commentato, è nella Figura R 6.3.

- Nella sua forma originaria ed essenziale, il box-plot è un grafico fondato sui quartili e che esplicita anche il valore minimo e quello massimo della variabile. Tali indicatori sono rappresentati – rispetto ad un conveniente asse – mediante una «scatola che si estende tra il primo ed il terzo quartile»: all'interno di tale scatola è evidenziata la mediana della distribuzione.

- Alle due estremità della scatola le osservazioni atipiche si evidenziano con appositi segni (per esempio, con •).
- Dal punto di vista operativo, si determinano i *capisaldi* della rappresentazione, cioè i seguenti valori: il valore minimo $min(X)$; il 1° quartile Q_1; la mediana Me; il 3° quartile Q_3; il valore massimo $max(X)$, nonché la differenza interquartile $IQR = Q_3 - Q_1$.
- Il disegno del box-plot avviene nel modo seguente (si veda la Figura R 6.3):

i) in un prefissato asse (verticale oppure orizzontale), la cui scala è quella della variabile in esame, si individuano delle barre in corrispondenza della mediana, del 1° e del 3° quartile, che poi si chiudono sino a formare una scatola. Le linee esterne che si dipartono da tale scatola si chiamano «baffi» (*whiskers*) per cui la rappresentazione è talvolta chiamata «*box-and-whiskers plot*»;

ii) si determinano delle linee verticali (oppure orizzontali) che escono dalla «scatola» nelle due direzioni opposte fino agli estremi, oppure sino a distanza $1.5\,IQR$, rispettivamente, dal primo e terzo quartile, costituendo quelle che si definiscono «cerniere» o «cardini» (*hinges*) .

L'utilità del box-plot deriva dal fatto che in una sola ed essenziale rappresentazione grafica si evidenziano tutti gli aspetti di una distribuzione di frequenza sin qui ritenuti importanti e che manifestano immediatamente la sintesi di una variabile statistica da molteplici punti di vista.

Infatti, il box-plot indica la *posizione* della variabile mediante la mediana e la *variabilità* mediante l'ampiezza della scatola (che è la differenza interquartile); inoltre, il collocamento della mediana all'interno della predetta scatola e rispetto agli estremi della distribuzione mostra anche l'eventuale *asimmetria* e la sua direzione.

Infine, trattandosi di una rappresentazione grafica *stilizzata*, il box-plot permette un immediato confronto tra più variabili o tra la stessa variabile esaminata in tempi, luoghi, occasioni differenti. Nel paragrafo R 6.3 sono illustrate molte ed importanti utilizzazioni dei box-plot, con esemplificazioni derivate dai data set.

6.6 Considerazioni finali

A conclusione dello studio sugli aspetti essenziali della distribuzione di una variabile X, dovrebbe essere emerso il ruolo essenziale della descrizione accurata di un fenomeno tramite la visualizzazione dei suoi aspetti peculiari. Ciò è avvenuto sia mediante la distribuzione di frequenza (nella sua formulazione originaria tramite diagramma a barre o istogramma, oppure con l'istogramma perequato, oppure resa essenziale con il box-plot) che mediante la proposta di numerosi indici descrittivi.

- La filosofia che ci ha accompagnato nell'analisi statistica delle osservazioni è stata quella di pervenire ad una fotografia della realtà per sintetizzare un fenomeno in rapporto a obiettivi predefiniti.

- L'ambito di tale studio è stato quello di una conoscenza completa e definitiva della popolazione, per cui *in questa fase* ha poco senso ogni domanda circa la possibilità di generalizzare i risultati trovati ad altri contesti, oppure la ricerca di forme teoriche che avrebbero migliori requisiti di semplicità rispetto ai dati reali.
- Un elemento che dovrebbe essere emerso dalle analisi precedenti è il fatto che lo studio di una distribuzione richiede una conoscenza preliminare della sua forma. In effetti, sin dalla introduzione della media aritmetica, quale sintesi per eccellenza della posizione della variabile X, è risultato ovvio che il «valore» di ogni indice statistico deriva strettamente da una misura della sua accuratezza e, quindi, dalla forma di una distribuzione, soprattutto dal rapporto tra il nucleo centrale dei valori e il peso dei valori estremi.

Per una variabile quantitativa X, quindi, assieme alla visualizzazione grafica (e qui istogramma perequato e box-plot possono essere utili soluzioni) suggeriamo di esplicitare sempre gli estremi, i decili, i quartili e la mediana, assieme alla media, lo scarto quadratico medio, e gli indici di forma prima introdotti secondo lo schema seguente:

Indici derivati dalla funzione di ripartizione						
$min(X)$	D_1	Q_1	Me	Q_3	D_9	$max(X)$

Indici derivati dalla distribuzione di frequenza			
μ	σ	γ	β

La conoscenza di tali informazioni offre una descrizione sufficientemente completa e, spesso, molto utile per uno studio esplorativo e riassuntivo della variabile in esame.

Ambiente R

Forma di una distribuzione di frequenza

R6.1 Asimmetria e curtosi di una distribuzione

Un primo ***indice di asimmetria*** normalizzato è basato cul confronto tra media e mediana. Supponiamo che sia già attiva la funzione varpop introdotta nel paragrafo 5.1. Esemplifichiamo tali calcoli per la variabile pesovivo nel data set camellia e per la variabile durata nel data set compact, rispettivamente:

```
R==> dati1=read.table("camellia.txt", header=TRUE);
R==> pesovivo=dati1[,4];

R==> dati2=read.table("compact.txt", header=TRUE);
R==> durata=dati2[,1];

R==> asym1=(mean(pesovivo)-median(pesovivo))/
          sqrt(varpop(pesovivo));
R==> asym2=(mean(durata)-median(durata))/sqrt(varpop(durata));
R==> asym1; asym2;
[1] 0.038023
[1] 0.13984
```

Quindi, la variabile pesovivo è quasi simmetrica mentre durata è asimmetrica positiva.

Gli indici più diffusi di ***asimmetria*** e ***curtosi*** sono definiti tramite le medie delle terze e quarte potenze delle variabili standardizzate, rispettivamente. Gli esempi, per le precedenti variabili, evidenziano come ottenere questi indici:

```
R==> zetapes=(pesovivo-mean(pesovivo))/sqrt(varpop(pesovivo));
R==> zetadur=(durata-mean(durata))/sqrt(varpop(durata));

R==> skewpes=mean(zetapes^3);    ### Asimmetria di pesovivo
R==> skewdur=mean(zetadur^3);    ### Asimmetria di durata

R==> curtpes=mean(zetapes^4);    ### Curtosi di pesovivo
R==> curtdur=mean(zetadur^4);    ### Curtosi di durata;
R==> skewpes; skewdur;
[1] 0.45071
[1] 1.8239
```

```
R==> curtpes; curtdur;
[1] 3.0097
[1] 9.336
```

L'asimmetria di pesovivo è modesta mentre più accentuata (e di segno positivo) è quella di durata (la distribuzione ha una evidente coda a destra). Invece, la curtosi di pesovivo è quasi esattamente pari a 3 (e questo denota una forma vicina a quella Gaussiana, una distribuzione teorica di cui ci occuperemo nel Capitolo 10) mentre la curtosi di durata è elevata (e ciò indica una distribuzione molto appuntita).

In R esiste anche la funzione scale() che standardizza il vettore o tutte le colonne della matrice su cui è applicata. Si vedano gli esempi seguenti:

```
R==> dati=read.table("camellia.txt", header=TRUE);
R==> pesovivo=dati[,4];
R==> zpes=scale(pesovivo); ### standardizza la variabile pesovivo
R==> round(mean(zpes)); ### arrotonda la media che è: -1.8511e-16
[1] 0
R==> var(zpes);  ### la varianza è una matrice, qui 1 x 1
     [,1]
[1,]    1
```

Senza opzioni, scale() sottrae la media e divide per lo scarto quadratico medio della variabile; ma è possibile sottrarre ai dati una costante qualsiasi e poi dividerli per un'altra costante, mediante idonee opzioni. In tal modo, si ottiene una generica trasformazione lineare della variabile.

R6.2 Istogramma perequato

La funzione grafica secondaria density(), con molte opzioni su cui non ci soffermiamo, genera in R le informazioni mediante le quali si può disegnare un ***istogramma perequato*** (ed anche sovrapporlo all'istogramma classico, eventualmente). Le istruzioni qui di seguito riportate sono quelle necessarie per riprodurre la Figura R 6.2 del testo nella quale si confrontano gli istogrammi perequati per la variabile peso nel data set studenti, distinta per genere.

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> peso=dati[,4];genere=dati[,2];
R==> pesom=peso[genere==0];  pesof=peso[genere==1];
R==> plot(density(pesom),main="",lwd=2,
          xlim=c(30,130),ylim=c(0,0.07));
R==> text(100,0.01,"Studenti",font=4);
R==> lines(density(pesof),lwd=2);
R==> text(40,0.055,"Studentesse",font=4);
```

Tali istruzioni mostrano che la funzione (primaria) plot() è utilizzata per il primo grafico relativo agli studenti e, poi, la funzione (secondaria) lines() per sovrapporvi quello relativo alle studentesse. La Figura R 6.1 evidenzia come sia netto lo spostamento verso valori più bassi della variabile peso delle studentesse e come la distribuzione del peso sia anche meno variabile e più appuntita per le donne. Si osservi che il grafico presentato è normalizzato (cioè l'area sotto ciascuna delle due curve è pari a 1) per cui i due istogrammi perequati sono confrontabili anche in presenza di numerosità molto differenti per i due generi. Infine, i valori che R stampa al di sotto dell'asse delle ascisse riportano su quanti punti il grafico è calcolato nonché un indicatore del grado di *smoothness* adottato nella perequazione.

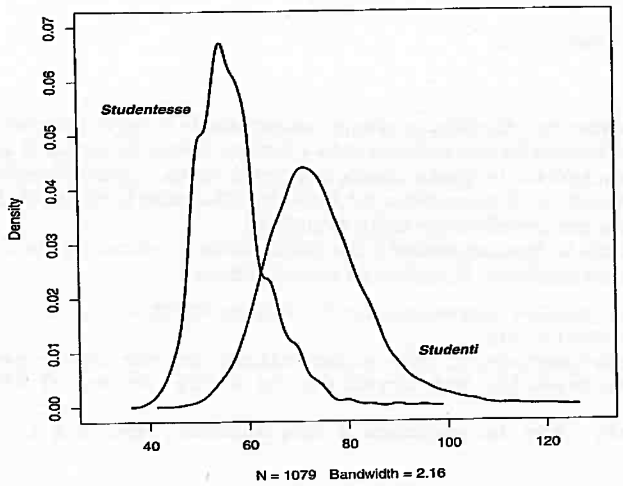

FIGURA R6.1. *Istogrammi perequati della variabile peso, per genere.*

Un'ulteriore possibilità (non descritta nel testo) è quella di aggiungere –mediante la funzione grafica secondaria rug(x)– dei segni verticali (*ticks*) sull'asse delle ascisse in corrispondenza delle modalità osservate, in modo da preservare la localizzazione dei dati originari, che viene invece mediata con l'istogramma perequato. La Figura R 6.2 mostra l'uso di tale funzione in aggiunta all'istogramma perequato della variabile durata nel data set compact, ed è ottenuta mediante le istruzioni seguenti:

```
R==> dati=read.table("compact.txt", header=TRUE);
R==> durata=dati[,1];
R==> plot(density(durata),type="l",lwd=2,
       main="Variabile Durata", ylab="Istogramma perequato");
R==> rug(durata,lwd=0.25,side=1);
```

L'opzione side=1 (oppure side=3) colloca i segni al di sotto delle ascisse (oppure in alto rispetto al grafico).

R6.3 Box-plot

Tra le analisi esplorative dei dati, la funzione boxplot() è una delle più versatili ed utili, dotata di numerose opzioni. Nella sua forma standard, effettua il ***box-plot*** delle modalità contenute nel vettore variabile ovvero nella variabile nomevariabile inclusa nel data set dati ovvero tanti box-plot per quante colonne sono presenti nella matrice mat, mediante una delle seguenti istruzioni (alternative tra loro):

```
R==> boxplot(variabile);                      ### 1° alternativa
R==> with(dati,boxplot(nomevariabile));       ### 2° alternativa
R==> boxplot(mat);                            ### 3° alternativa
```

Nella Figura R 6.3 è evidenziato un box-plot di tipo standard costruito sulla variabile peso per il data set studenti e, sul lato sinistro, sono descritti gli indicatori che definiscono

```
R==> curtpes; curtdur;
[1] 3.0097
[1] 9.336
```

L'asimmetria di pesovivo è modesta mentre più accentuata (e di segno positivo) è quella di durata (la distribuzione ha una evidente coda a destra). Invece, la curtosi di pesovivo è quasi esattamente pari a 3 (e questo denota una forma vicina a quella Gaussiana, una distribuzione teorica di cui ci occuperemo nel Capitolo 10) mentre la curtosi di durata è elevata (e ciò indica una distribuzione molto appuntita).

In R esiste anche la funzione scale() che standardizza il vettore o tutte le colonne della matrice su cui è applicata. Si vedano gli esempi seguenti:

```
R==> dati=read.table("camellia.txt", header=TRUE);
R==> pesovivo=dati[,4];
R==> zpes=scale(pesovivo); ### standardizza la variabile pesovivo
R==> round(mean(zpes)); ### arrotonda la media che è: -1.8511e-16
[1] 0
R==> var(zpes);  ### la varianza è una matrice, qui 1 x 1
     [,1]
[1,]    1
```

Senza opzioni, scale() sottrae la media e divide per lo scarto quadratico medio della variabile; ma è possibile sottrarre ai dati una costante qualsiasi e poi dividerli per un'altra costante, mediante idonee opzioni. In tal modo, si ottiene una generica trasformazione lineare della variabile.

R6.2 Istogramma perequato

La funzione grafica secondaria density(), con molte opzioni su cui non ci soffermiamo, genera in R le informazioni mediante le quali si può disegnare un ***istogramma perequato*** (ed anche sovrapporlo all'istogramma classico, eventualmente). Le istruzioni qui di seguito riportate sono quelle necessarie per riprodurre la Figura R 6.2 del testo nella quale si confrontano gli istogrammi perequati per la variabile peso nel data set studenti, distinta per genere.

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> peso=dati[,4];genere=dati[,2];
R==> pesom=peso[genere==0];  pesof=peso[genere==1];
R==> plot(density(pesom),main="",lwd=2,
          xlim=c(30,130),ylim=c(0,0.07));
R==> text(100,0.01,"Studenti",font=4);
R==> lines(density(pesof),lwd=2);
R==> text(40,0.055,"Studentesse",font=4);
```

Tali istruzioni mostrano che la funzione (primaria) plot() è utilizzata per il primo grafico relativo agli studenti e, poi, la funzione (secondaria) lines() per sovrapporvi quello relativo alle studentesse. La Figura R 6.1 evidenzia come sia netto lo spostamento verso valori più bassi della variabile peso delle studentesse e come la distribuzione del peso sia anche meno variabile e più appuntita per le donne. Si osservi che il grafico presentato è normalizzato (cioè l'area sotto ciascuna delle due curve è pari a 1) per cui i due istogrammi perequati sono confrontabili anche in presenza di numerosità molto differenti per i due generi. Infine, i valori che R stampa al di sotto dell'asse delle ascisse riportano su quanti punti il grafico è calcolato nonché un indicatore del grado di *smoothness* adottato nella perequazione.

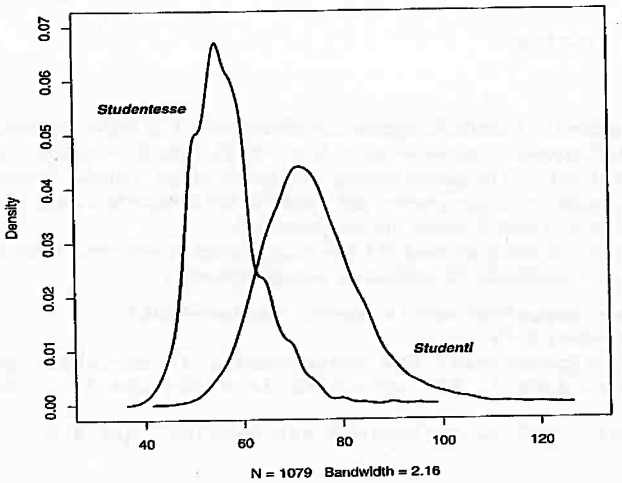

FIGURA R6.1. *Istogrammi perequati della variabile peso, per genere.*

Un'ulteriore possibilità (non descritta nel testo) è quella di aggiungere –mediante la funzione grafica secondaria rug(x)– dei segni verticali (*ticks*) sull'asse delle ascisse in corrispondenza delle modalità osservate, in modo da preservare la localizzazione dei dati originari, che viene invece mediata con l'istogramma perequato. La Figura R 6.2 mostra l'uso di tale funzione in aggiunta all'istogramma perequato della variabile durata nel data set compact, ed è ottenuta mediante le istruzioni seguenti:

```
R==> dati=read.table("compact.txt", header=TRUE);
R==> durata=dati[,1];
R==> plot(density(durata),type="l",lwd=2,
      main="Variabile Durata", ylab="Istogramma perequato");
R==> rug(durata,lwd=0.25,side=1);
```

L'opzione side=1 (oppure side=3) colloca i segni al di sotto delle ascisse (oppure in alto rispetto al grafico).

R6.3 Box-plot

Tra le analisi esplorative dei dati, la funzione boxplot() è una delle più versatili ed utili, dotata di numerose opzioni. Nella sua forma standard, effettua il ***box-plot*** delle modalità contenute nel vettore variabile ovvero nella variabile nomevariabile inclusa nel data set dati ovvero tanti box-plot per quante colonne sono presenti nella matrice mat, mediante una delle seguenti istruzioni (alternative tra loro):

```
R==> boxplot(variabile);                    ### 1° alternativa
R==> with(dati,boxplot(nomevariabile));     ### 2° alternativa
R==> boxplot(mat);                          ### 3° alternativa
```

Nella Figura R 6.3 è evidenziato un box-plot di tipo standard costruito sulla variabile peso per il data set studenti e, sul lato sinistro, sono descritti gli indicatori che definiscono

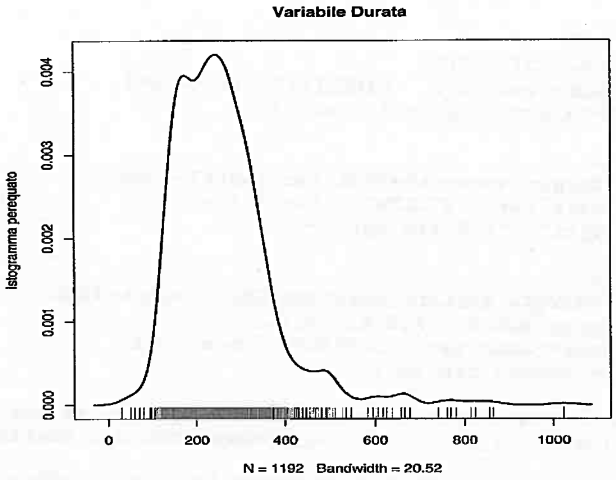

FIGURA R6.2. *Uso della funzione rug per la variabile durata nel data set compact.*

la scatola, i "baffi", i valori anomali e la collocazione degli estremi. La figura è stata ottenuta con le seguenti istruzioni:

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> x=dati[,4];
R==> boxplot(x,ylab="Peso (in Kg)",col=gray(0.6));
R==> testo=c("minimo","cardine inferiore","1° quartile",
        "mediana","3° quartile","cardine superiore",
        "valori anomali","massimo");
R==> ascissetesto=rep(1.275,length(testo));
R==> ordinatetesto=c(min(x),min(x)+2,
        quantile(x,c(0.25,0.50,0.75),names=FALSE),
        92,max(x)-10,max(x));
R==> text(ascissetesto,ordinatetesto,testo,cex=0.7,font=4);
```

Tra le numerose opzioni che migliorano la capacità rappresentativa del box-plot, citiamo le seguenti:

- `varwidth=TRUE` consente di rappresentare l'ampiezza della scatola in modo proporzionale alla radice quadrata del numero di modalità (questa opzione è preferibile nel presentare box-plot di variabili relative a sottogruppi di numerosità differenti).
- `notch=TRUE` crea un intaglio (*notch*) al centro della scatola, costruito in modo tale che se due intagli di differenti box-plot non si sovrappongono, verosimilmente, le due mediane differiscono.
- `col=...` consente di specificare il vettore (di caratteri) che definisce i colori per la scatola centrale del (o dei) box-plot. Sfumature graduate del colore – se scelte accuratamente – possono aiutare, talvolta, nella interpretazione dei dati.
- `horizontal=TRUE` consente di disegnare il (i) box-plot in modo orizzontale.

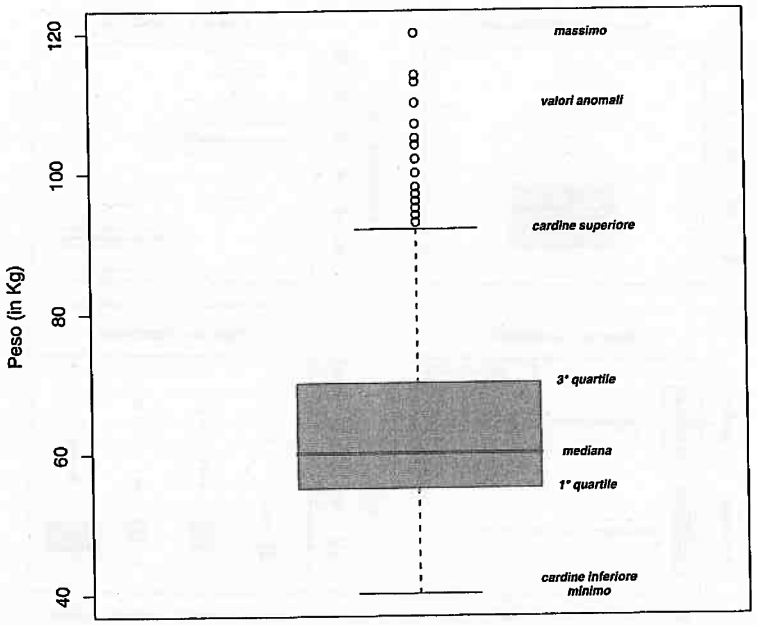

FIGURA R6.3. *Box-plot della variabile peso per il data set studenti.*

- `add=TRUE` consente di aggiungere il (i) box-plot ad un grafico già esistente.
- Se l'argomento della funzione `boxplot` è una formula (secondo il significato di `R` , come sarà illustrato brevemente nel Capitolo 18) del tipo `variabile~fattore` (dove `variabile` è quantitativa mentre `fattore` è qualitativa), allora la funzione disegna i box-plot condizionati al valore specificato da `fattore`: tale opzione è molto utile per verificare se la distribuzione di `variabile` si modifica al variare delle modalità (=*levels*) di `fattore` (per esempio, al variare del tempo, del genere, dei gruppi, del territorio, delle circostanze, e così via).

Presentiamo, ora, nella Figura `R` 6.4 alcune applicazioni della funzione `boxplot()`, per differenti data set, allo scopo di evidenziare l'effetto grafico di alcune opzioni, prescelte di volta in volta. In particolare, dopo il richiamo dei differenti data set, sono anche elencati i comandi per creare fattori in modo opportuno. Più precisamente: due soli livelli, nel caso del `genere` per il data set `studenti`; quattro livelli, definiti mediante la funzione `cut()` applicata su `anno` per definire la variabile `durata` rispetto al `periodo` (nuovo fattore così creato) nel data set `compact`. Questi accorgimenti consentono di disegnare box-plot distinti per differenti sottogruppi e rendere agevole il confronto sulla medesima scala. Infine, si osservi l'opzione `mar()` tramite la quale si estendono i margini inferiori e superiori di ciascun riquadro per migliorarne la visualizzazione, e l'opportunità della funzione `legend()` aggiunta nel terzo box-plot per definire l'unità di misura sull'asse.

```
### Lettura preliminare dei dati e trasformazione in fattori

R==> dati1=read.table("prezzolibri.txt", header=TRUE);
```

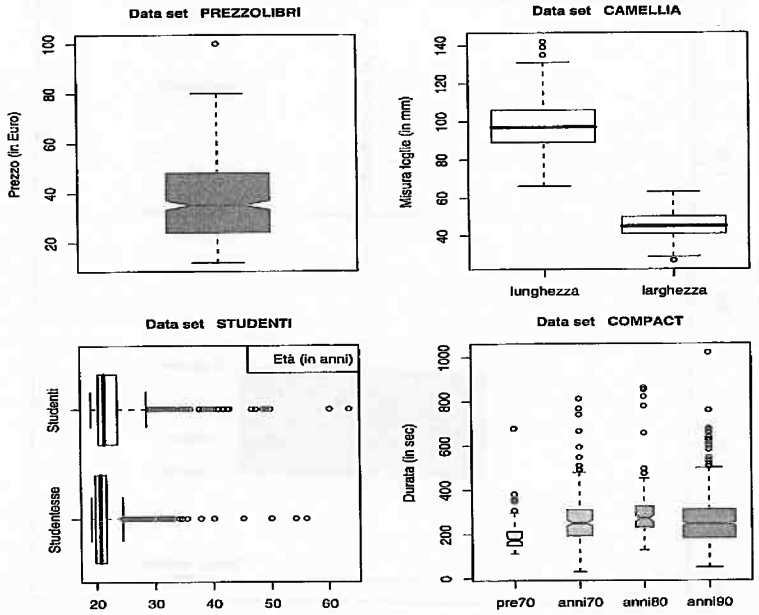

FIGURA R6.4. *Tipologie di box-plot, ottenuti con differenti opzioni.*

```
R==> dati2=read.table("camellia.txt", header=TRUE);

R==> dati3=read.table("studenti.txt", header=TRUE);
R==> eta=dati3[,3]; genere=dati3[,2];
R==> facgen=as.factor(ifelse(genere==0,"Studenti","Studentesse"));

R==> dati4=read.table("compact.txt", header=TRUE);
R==> durata=dati4[,1]; anno=dati4[,2];
R==> periodo=cut(anno,breaks=c(1958,1970,1980,1990,1996),
       labels=c("pre70","anni70","anni80","anni90"),
       include.lowest = TRUE,ordered_result =TRUE);

### Esempio di quattro differenti tipologie di box-plot

R==> par(mfrow=c(2,2));        # suddividi area grafica
                               # in 4 regioni
R==> par(mar=c(2,4,2,2)+0.1); # amplia margini
                               # inferiore e superiore

### Grafico n.1
R==> boxplot(dati1[,1], notch=TRUE,col=gray(0.4));
R==> title(main="Data set   PREZZOLIBRI",cex.main=1,
       ylab="Prezzo (in Euro)");
```



```
### Grafico n.2
R==> boxplot(dati2[,2:3]);
R==> title(main="Data set   CAMELLIA",cex.main=1,
       ylab="Misura foglie (in mm)");

# Grafico n.3
boxplot(eta~facgen,varwidth=TRUE,horizontal=TRUE);
title(main="Data set   STUDENTI",cex.main=1);
legend("topright", "Età (in anni)");

### Grafico n.4
R==> boxplot(durata~periodo,varwidth=TRUE, notch=TRUE,
       col=gray(c(0.8,0.7,0.6,0.5)));
R==> title(main="Data set   COMPACT",cex.main=1,
       ylab="Durata (in sec)");

R==> par(mar=c(5,4,4,2)+0.1) # ripristina opzioni margini
R==> par(mfrow=c(1,1))       # ripristina area del grafico
```

Nella letteratura esistono molte varianti dei box-plot e svariate utilizzazioni combinate: una di esse, resa disponibile da Verzani (2005) e presente all'interno del *package* `UsingR`, è la funzione `simple.hist.and.boxplot()` che ci pare molto efficace per essere utilizzata nelle analisi esplorative dei dati. Infatti, in una sola rappresentazione grafica, essa include l'istogramma, il box-plot e l'elencazione ordinata delle modalità mediante `rug()`.

Le istruzioni seguenti (che generano la Figura R 6.5) richiamano anzitutto il *package* `UsingR` e poi la funzione `simple.hist.and.boxplot()` applicata alla variabile peso nel data set `studenti`:

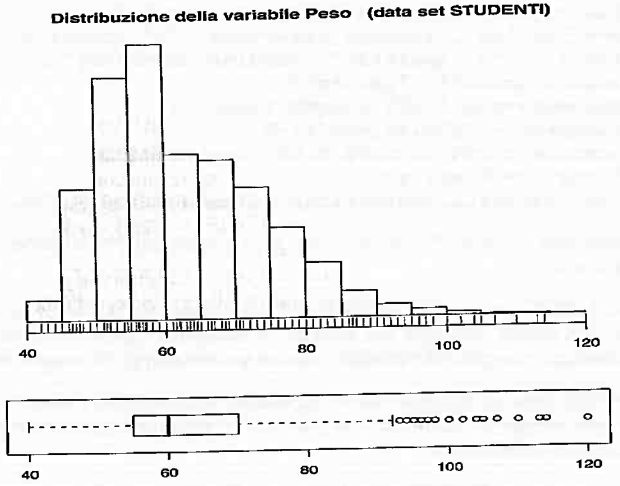

FIGURA R6.5. *Grafico congiunto di istogramma, boxplot e dati originari.*

```
R==> dati=read.table("studenti2.txt", header=TRUE); peso=dati[,4];
R==> require("UsingR");
R==> simple.hist.and.boxplot(peso,
     main="Distribuzione della variabile
     Peso (data set STUDENTI)");
```

R6.4 Trasformazioni delle variabili

È importante applicare metodi statistici su variabili che siano di forma simmetrica ed unimodali, per motivi che diverranno più evidenti quando si passerà allo studio dell'Inferenza statistica (in effetti, è preferibile che la distribuzione osservata approssimi quella della variabile casuale Normale, che studieremo nel Capitolo 10). Qui ci limitiamo ad evidenziare se e come una trasformazione dei dati produca l'effetto ricercato, accentuando l'aspetto grafico di tale ricerca.

Esiste una vasta famiglia di trasformazioni (detta di Box-Cox dal nome dei due grandi statistici che la studiarono nel 1964), che riprenderemo nel paragrafo 13.2 e che include la radice quadrata, la trasformazione identica, l'inversa dei dati, la trasformazione logaritmica, e così via. Per una serie di *modalità non-negative* $\{x_1, x_2, \ldots, x_n\}$, la ***trasformazione di Box-Cox*** è definita in funzione di un parametro reale λ da:

$$x_i(\lambda) = \begin{cases} \dfrac{x_i^\lambda - 1}{\lambda}, & \text{per } \lambda \neq 0; \\ \log(x_i), & \text{per } \lambda = 0; \end{cases} \qquad i = 1, 2, \ldots, n.$$

A tal fine, per agevolare i calcoli successivi, abbiamo codificato una funzione
`trasf(x,lambda)`
nel modo seguente:

```
trasf<-function(x,lambda){
if(lambda==0) log(x) else(x^lambda-1)/lambda
}
```

nella quale la funzione `if...() else...()` significa che se (=*if*) $\lambda = 0$ la funzione calcola il logaritmo, altrimenti (=*else*) calcola $\frac{x_i^\lambda - 1}{\lambda}$, come nella definizione precedente.

In effetti, se $\lambda = -1, -1/2, 0, 1/3\,1/2, 1$, la trasformazione consiste – rispettivamente – nello studiare l'inverso dei dati (se $\lambda = -1$), la radice dell'inverso (se $\lambda = -1/2$), il logaritmo (se $\lambda = 0$), la radice cubica (se $\lambda = 1/3 \simeq 0.333$), la radice quadrata (se $\lambda = 1/2$) ovvero non trasformarli affatto (se $\lambda = 1$).

La scelta può avvenire in modo rapido confrontando l'istogramma perequato per questi sei casi ed individuando la trasformazione che meglio raggiunge l'obiettivo, come ora esemplifichiamo nella Figura R 6.6 per la variabile durata nel data set compact, mediante le seguenti istruzioni:

```
R==> dati=read.table("compact.txt", header=TRUE);
R==> durata=dati[,1];

R==> par(mfrow=c(3,2));          # suddividi l'area grafica
                                 # in 6 regioni
R==> par(mar=c(2,4,2,2)+0.1); # amplia margini
                                 # inferiore e superiore

# Grafico n.1
```


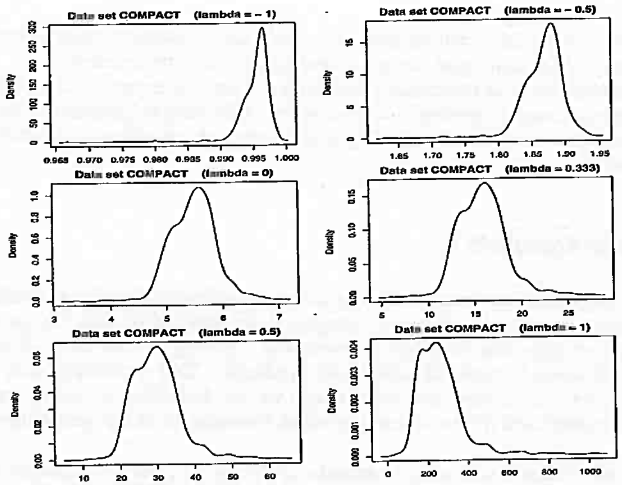

FIGURA R6.6. *Istogrammi perequati della variabile durata e sue trasformazioni, nel data set compact.*

```
R==> plot(density(trasf(durata,-1)),xlab="",
       main="Data set COMPACT    (lambda = - 1)");
# Grafico n.2
R==> plot(density(trasf(durata,-0.5)),xlab="",
       main="Data set COMPACT    (lambda = - 0.5)");
# Grafico n.3
R==> plot(density(trasf(durata,0)),xlab="",
       main="Data set COMPACT    (lambda = 0)");
# Grafico n.4
R==> plot(density(trasf(durata,1/3)),xlab="",
       main="Data set COMPACT    (lambda = 0.333)");
# Grafico n.5
R==> plot(density(trasf(durata,0.5)),xlab="",
       main="Data set COMPACT    (lambda = 0.5)");
# Grafico n.6
R==> plot(density(durata),xlab="",
       main="Data set COMPACT    (lambda = 1)");

R==> par(mar=c(5,4,4,2)+0.1); # ripristina opzioni margini
R==> par(mfrow=c(1,1));       # ripristina l'area del grafico
```

La Figura R 6.6 mostra come la variabile durata sia asimmetrica positiva (la coda a destra quando $\lambda = 1$ è ben evidente) e tale forma si modifica mediante la trasformazione proposta raggiungendo una sufficiente simmetria quando $0.5 \leq \lambda \leq 0.333$. Questo caso mostra come la trasformazione logaritmica (che, in effetti, è quella più diffusa nelle applicazioni) non sia sempre la più idonea anche perché, verosimilmente, la variabile considerata è bimodale (quindi, consiste del miscuglio di più distribuzioni).

Queste analisi esplorative hanno lo scopo di indurre nel lettore il desiderio di sviluppare un'analisi critica sui propri dati non assumendo come valide né le proposte né le assunzioni

codificate nei libri di testo: una conoscenza del problema in esame (e degli obiettivi della ricerca) deve sempre affiancare ogni corretta analisi statistica preliminare.

Infine, precisiamo che nelle istruzioni precedenti abbiamo omesso di utilizzare funzioni di R più complesse per una immediata comprensione delle singole operazioni ma più efficienti dal punto di vista computazionale. Il paragrafo successivo affronta preliminarmente questa problematica.

R6.5 Operazioni ripetute in R

Nelle analisi statistiche avviene spesso che occorre replicare operazioni standard su un gruppo di variabili (=colonne) o di unità (=righe) della matrice di dati. Un primo modo è quello di creare un'apposita funzione (=*function*) che esegue una serie di operazioni semplicemente indicando le variabili sulle quali applicarle. Tali procedure sono state più volte utilizzate sinora e lo saranno ancora nel seguito: per le finalità di queste note non ci addentriamo nei dettagli ma il lettore comprenderà l'essenziale di tali procedure, di volta in volta.

In aggiunta alle `function()` appositamente costruite, si possono replicare le singole istruzioni in un ciclo oppure utilizzare delle altre funzioni tipiche dell'ambiente R (più complesse ma spesso più rapide).

Sul primo versante, l'operazione più semplice (comune a tutti i linguaggi di programmazione) è quella del *loop* (=ciclo) mediante il quale si replica la stessa operazione un certo numero di volte. Per esempio, per calcolare e visualizzare la media delle variabili numeriche presenti nel data set `camellia` (tali variabili sono nelle colonne da 2 a 5) si possono eseguire le istruzioni:

```
R==> dati=read.table("camellia.txt", header=TRUE);

R==> for(j in 2:5){
        print(mean(dati[,j]))
     }
[1] 97.565
[1] 45.718
[1] 9985.2
[1] 3961.5
```

Nello specifico, al variare dell'indice j nell'insieme $j = 2, 3, 4, 5$, l'istruzione `for` ripete tutto quello che è inserito tra le parentesi $\{\ldots\}$ facendo assumere all'indice j, sequenzialmente, i valori $j = 2$, $j = 3$, $j = 4$, $j = 5$. Nel caso evidenziato, calcola ed invia sullo schermo la media delle variabili poste nelle colonne $2, 3, 4, 5$ del data frame `dati`. Ovviamente, tale scelta ha uno scopo esemplificativo poiché si è già mostrato in precedenza come ottenere in modo rapido la media di tutte le colonne di un *data frame*. Ebbene, la funzione `for()` è più generale di quanto indicato perché può replicare le operazioni rispetto agli elementi di qualsiasi lista predefinita (non solo numeri, quindi) ma il suo principale difetto è la lentezza nelle elaborazioni. Esistono in R altre funzioni per eseguire cicli (come `repeat`, `while`, etc.) ma nel seguito ci limiteremo ad applicare solo la funzione `for()`.

Tuttavia, se dovessimo ripetere un'operazione per un milione di volte, come accade in alcuni esperimenti di simulazione (uno dei quali sarà proposto nel paragrafo 11.2), i tempi possono diventare inaccettabili, per cui tranne casi semplici (come, per motivi didattici, avviene in queste note), una efficiente programmazione presuppone un principio molto semplice: *occorre assolutamente evitare i loop*. In R questo è possibile utilizzando il calcolo matriciale oppure particolari funzioni tra le quali ci limitiamo ad esaminare `apply()` e `tapply()`.

Nel seguito, forniremo poche esemplificazioni facendo riferimento al data set `studenti`; poi, nelle pagine seguenti, tali funzioni saranno utilizzate quando opportuno.

`apply()` La sintassi di questa funzione, che si applica ad una *matrice*, è la seguente:

```
R==> apply(matrice, codice, funzione);
```

dove il codice vale 1 se la funzione da applicare opera per righe e vale 2 se opera per colonne.

```
R==> dati=read.table("studenti2.txt", header=TRUE);
### calcola la lunghezza di tutte le variabili
R==> with(data=dati, apply(dati,2,length));
      anno     genere        eta       peso    altezza
      2759       2759       2759       2759       2759
    dipsci     dipcla     diptec     dipalt componenti
      2759       2759       2759       2759       2759
  occhiali      fumo     diploma
      2759       2759       2759

### calcola la media delle variabili eta, peso, altezza
R==> with(data=dati, apply(dati[,3:5],2,mean)) ;
     eta     peso altezza
 21.851  63.381 169.012

###  calcola la mediana delle variabili eta, peso, altezza
R==> with(data=dati, apply(dati[,3:5],2,median)) ;
     eta     peso altezza
    20.8     60.0    168.0
```

`tapply()` La sintassi di questa funzione, che si applica ad un *data frame*, è la seguente:

```
R==> tapply(dataframe, rispetto-a-cosa, funzione);
```

La sua maggior utilità è quando si devono calcolare indici o funzioni per sottogruppi oppure per diversi livelli di un fattore, come nell'esempio seguente:

```
### calcola età media rispetto al genere
R==> with(data=dati, tapply(eta,genere,mean));
    0     1
22.77 21.26

### calcola età mediana rispetto al diploma
R==> with(data=dati, tapply(eta,diploma,median));
 alt  cla  sci  tec
20.7 20.7 20.8 21.1
```

In tutti i casi considerati, è possibile includere l'opzione `na.rm=TRUE` che evita di applicare le funzioni considerate alle unità con dati mancanti. Va segnalato che numerose altre funzioni di questo tipo sono utili per casi più complessi.

Capitolo settimo

Distribuzioni statistiche multiple

7.1 Introduzione

Una delle finalità più comuni della raccolta di dati è la ricerca di relazioni di causa-effetto tra fenomeni, allo scopo di interpretare, prevedere, simulare, controllare. Per questo tipo di studi è essenziale analizzare il contemporaneo presentarsi delle modalità di più variabili, anziché limitarsi allo studio delle singole distribuzioni.

Quando su ogni unità statistica, appartenente ad una determinata popolazione, si rilevano due, tre, ..., più variabili (qualitative e/o quantitative) si parla di distribuzione doppia, tripla, ..., multipla (o multivariata).

Lo studio diventa necessariamente più complesso e coinvolge numerosi aspetti che non possono essere affrontati in questo volume. Qui ci limiteremo alla discussione delle variabili doppie e della misura del legame tra le loro componenti sia per variabili qualitative (connessione) che per variabili quantitative (correlazione). Quindi, accenneremo ad alcune problematiche generate dalle osservazioni multiple al fine di pervenire ad una corretta interpretazione del coefficiente di correlazione lineare.

7.2 Distribuzioni doppie di frequenza

Quando le variabili oggetto di studio sono due è comune indicare la variabile doppia con (X, Y) precisando che X è la prima componente e Y è la seconda componente.

- Una variabile doppia è *quantitativa* se entrambe le componenti sono quantitative. Per esempio, su ogni studente si rilevano (Peso, Altezza); in ogni Comune si rilevano (Nati, Residenti); di ogni oggetto fisico di un certo collettivo si rilevano (Temperatura, Umidità), etc. In tali casi, le variabili componenti potrebbero essere entrambe discrete, continue o delle due tipologie.
- Una variabile doppia è *qualitativa* se entrambe le componenti sono qualitative. Per esempio, su ogni studente si rilevano (Sesso, Colore dei capelli); di ogni auto si rilevano

(Colore della carrozzeria, Professione del proprietario); di ogni residente in Italia si rilevano (Stato civile, Regione di nascita), etc.

o Una variabile doppia è *mista* se una componente è qualitativa e l'altra quantitativa. Per esempio, su ogni studente si rilevano (Colore degli occhi, Peso); per ogni Comune si rilevano (Altezza sul livello del mare, Attività prevalente della popolazione attiva); di ogni frutto che si vende al mercato si rilevano (Colore della buccia, Percentuale di vitamina A), etc.

Se la rilevazione di una variabile doppia (X, Y) sulle n unità statistiche consiste nella collezione delle coppie ordinate:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_n, y_n),$$

ottenute rilevando, ordinatamente, su ciascuna delle n unità statistiche sia la variabile X che la variabile Y, si ha una *serie.*

Se, invece, le frequenze di tutte le coppie di modalità (x_i, y_j) sono disponibili in una distribuzione (doppia) di frequenze si ha una *seriazione.*

Nei casi in cui la distribuzione doppia consiste di molte unità statistiche, è opportuno procedere ad un raggruppamento dei dati considerando, in analogia al caso di una sola variabile, la distribuzione doppia delle frequenze con le quali ciascuna coppia di modalità si presenta nel collettivo in esame.

Supponiamo che la variabile X assuma le k modalità distinte: $x_1, x_2, \ldots, x_k$ e che la variabile Y assuma le h modalità distinte: $y_1, y_2, \ldots, y_h$.

Nel seguito indicheremo con n_{ij} la ***frequenza assoluta*** (cioè il numero delle volte) con la quale la coppia di valori (x_i, y_j) si presenta, cioè la frequenza con la quale su di una unità statistica la variabile X assume il valore x_i e *contemporaneamente* la variabile Y assume il valore y_j.

Analogamente, la ***frequenza relativa*** è la proporzione dei casi in cui, su una popolazione di n unità statistiche, la coppia di valori (x_i, y_j) si presenta, quindi: $f_{ij} = \dfrac{n_{ij}}{n}$, per ogni $i = 1, 2, \ldots, k$ e per ogni $j = 1, 2, \ldots, h$.

Per organizzare tali informazioni in forma tabellare, si utilizza la *tabella a doppia entrata* che esplicita sia le modalità delle due variabili che le rispettive frequenze assolute.

$X \backslash\backslash Y$	y_1	y_2	...	y_j	...	y_h	Totali
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1h}	$n_{1.}$
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2h}	$n_{2.}$
...	...	...	...	...	...	...	...
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ih}	$n_{i.}$
...	...	...	...	...	...	...	...
x_k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{kh}	$n_{k.}$
Totali	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.h}$	n

Le frequenze poste ai due margini della tabella doppia sono ottenute per addizione per riga e colonna rispettivamente. Esse sono dette ***frequenze marginali*** ed esplicitano la frequenza della modalità di una variabile senza tener conto delle modalità dell'altra.

o Dal punto di vista mnemonico, si osservi che le frequenze doppie n_{ij} della variabile (X, Y) possiedono due indici di cui il primo riguarda la modalità della X e il secondo quello della Y. Inoltre, la frequenza marginale $n_{i.}$ riguarda le frequenze della modalità i-esima della prima componente e il (.) ricorda che sono state sommate tutte le frequenze corrispondenti alla seconda componente, cioè la variabile Y. Similmente, la frequenza marginale $n_{.j}$ riguarda la frequenza della modalità j-esima della seconda componente e il (.) ricorda che sono state sommate tutte le frequenze corrispondenti alla prima componente, cioè la variabile X.

Per esempio, $n_{2.}$ è la frequenza della modalità x_2 per la variabile X senza tenere conto dei valori assunti da Y; quindi, $n_{2.}$ è il numero di unità statistiche che presentano la modalità x_2 per la variabile X a prescindere dai valori assunti per la variabile Y.

Similmente, $n_{.1}$ è la frequenza della modalità y_1 per la variabile Y senza tenere conto dei valori assunti da X; quindi, $n_{.1}$ è il numero di unità statistiche che presentano la modalità y_1 per la variabile Y a prescindere dai valori assunti per la variabile X.

Sul piano formale, frequenze doppie e frequenze marginali obbediscono ai seguenti vincoli:

$$\sum_{i=1}^{k}\sum_{j=1}^{h} n_{ij} = n; \qquad \sum_{i=1}^{k} n_{i.} = \sum_{j=1}^{h} n_{.j} = n$$

Tutte le precedenti quantità se divise per n producono le corrispondenti frequenze relative (doppie e marginali) ed una tabella esattamente analoga alla precedente, con il totale generale uguale a 1.

o Nella letteratura statistica italiana, talvolta, una tabella che contiene una distribuzione doppia di frequenze si definisce di *correlazione* se X e Y sono caratteri quantitativi, di *contingenza* se entrambi i caratteri sono qualitativi, di *associazione* se si prescinde dalla natura delle due variabili. Tale terminologia, che riportiamo per completezza, non sarà utilizzata in questo testo e non è comunque accolta da tutti gli Autori.

La discussione sin qui sviluppata riesce più semplice esaminando in dettaglio qualche esemplificazione, distinguendo i casi di una variabile doppia discreta, di una variabile doppia continua, di una variabile doppia mista (qualitativa e quantitativa).

Esempio 7.1

Raccogliendo informazioni su $n = 1000$ nuclei familiari, si intende studiare la variabile doppia (X, Y) che ha entrambe le componenti discrete definite da:

X = numero di auto possedute dalla famiglia;

Y = numero di componenti della famiglia.

Tali informazioni sono evidenziate nella seguente tabella delle frequenze assolute:

$X//Y$	1	2	3	4	5	*Totali*
0	10	20	20	150	50	250
1	85	85	330	50	50	600
2	5	85	10	0	0	100
3	0	10	40	0	0	50
Totali	100	200	400	200	100	1000

Il significato della tabella dovrebbe essere abbastanza evidente: per esempio, tra le 1000 famiglie intervistate, ve ne sono 330 che sono costituite da 3 componenti e *contemporaneamente* possiedono 1 sola auto; invece, 5 famiglie formate da 1 solo componente possiedono 2 auto, e così via. Se tutte le frequenze assolute della precedente tabella si dividono per il totale delle frequenze (in questo caso, per 1000) si ottiene una tabella analoga, detta delle frequenze relative (utile per confronti tra situazioni differenziate).

Può essere interessante (per esempio, per finalità di marketing) studiare la distribuzione delle auto possedute dalle famiglie che hanno un prefissato numero di componenti: per esempio, la distribuzione del numero di auto per le famiglie con 2 componenti. Evidentemente, dalla tabella precedente, si tratta di studiare la variabile X per la sola colonna relativa a $(Y = 2)$, ovvero, come si dice, «X condizionatamente al valore di $(Y = 2)$». In tal caso, si scrive $X|(Y = 2)$. La tabella seguente mostra, quindi, la distribuzione di frequenza della variabile $X|(Y = 2)$ con le rispettive frequenze assolute e relative.

$X\|(Y=2)$	Frequenze assolute	Frequenze relative
0	20	0.100
1	85	0.425
2	85	0.425
3	10	0.050
Totali	200	1.000

Quindi, una frazione di 0.425 (il 42.5%) delle famiglie con 2 componenti possiede 1 auto e una identica frazione ne possiede 2; invece, tra tali famiglie solo il 10% non possiede auto.

Poiché $X|(Y = 2)$ è una variabile discreta semplice (cioè univariata), per essa si possono calcolare tutti gli indici statistici studiati nei capitoli precedenti; per esempio, la sua media è ottenuta da:

$$\mu_{(X|Y=2)} = \frac{0 \times 20 + 1 \times 85 + 2 \times 85 + 3 \times 10}{200} = 1.425$$

Se si ripete tale calcolo per il numero medio di auto (X) condizionatamente al numero medio di componenti (Y), per tutti i valori disponibili di Y, si ottengono i seguenti risultati:

Variabili condizionate	Medie condizionate
$X\|(Y=1)$	0.950
$X\|(Y=2)$	1.425
$X\|(Y=3)$	1.175
$X\|(Y=4)$	0.250
$X\|(Y=5)$	0.500

Il risultato mostra che – nella popolazione considerata – il numero medio di auto cresce sino a 2 componenti per poi diminuire.

Se, invece di considerare le colonne della tabella a doppia entrata, si considerano le righe si ottengono le variabili Y condizionate ai valori di X, ovvero il numero di componenti condizionato ad un prefissato numero di auto.

□

Nel caso di variabile doppia quantitativa la rappresentazione grafica più naturale è quella del piano cartesiano ove si fa corrispondere ogni punto alla coppia ordinata di valori (x_i, y_i) rilevata sull'i-esima unità statistica, per $i = 1, 2, \ldots, n$. Tale rappresentazione viene chiamata «*nuvola di punti*» (in inglese, *scatter plot*).

Esempio 7.2

Dal data set STUDENTI, esaminiamo per gli $n = 2759$ studenti la variabile doppia (X, Y), dove: X = *Altezza* (in cm) e Y = *Peso* (in kg). Essa consiste nella elencazione: $(160, 61); (163, 55); (167, 57); \ldots; (177, 60); (175, 90)$; nella quale ogni coppia di valori indica, ordinatamente, l'altezza e il peso del $1°, 2°, \ldots,$ n-esimo studente.

Il grafico seguente rappresenta tali coppie di valori facendo corrispondere a ciascuna di esse un punto sul piano cartesiano, per il quale l'ascissa è la modalità assunta da X e l'ordinata quella assunta da Y nella unità statistica.

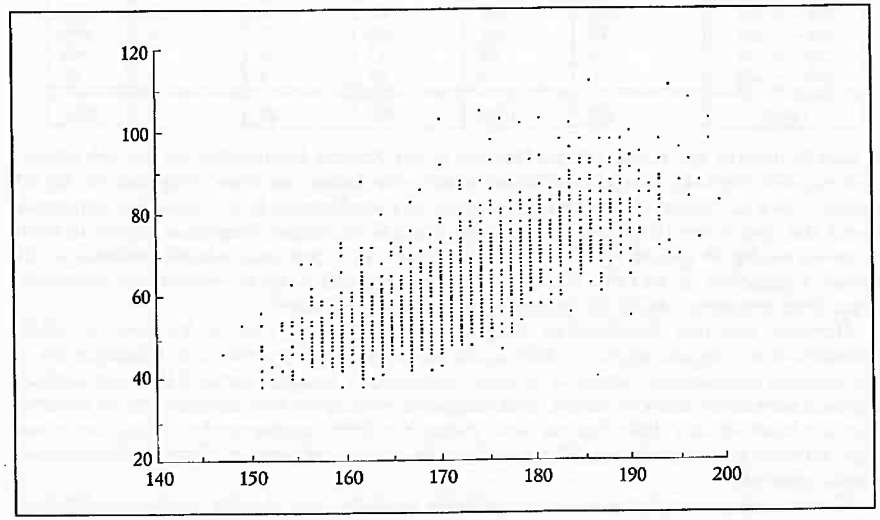

FIGURA 7.1. Variabile «Peso» rispetto a variabile «Altezza» per il data set STUDENTI.

La nuvola di punti evidenzia una notevole dispersione nei dati, anche se è evidente un orientamento ben definito nella configurazione dei punti. Si vede che, tendenzialmente, le persone alte sono più pesanti di quelle basse, e viceversa. Esiste cioè una relazione tra X e Y che non può essere espressa da una funzione matematica; in caso contrario, tutte le

persone alte cm 170, per esempio, dovrebbero avere lo stesso peso, il che è contraddetto dai dati, oltre che dall'esperienza di ciascuno.

□

Esempio 7.3

Negli esempi dei capitoli precedenti, sono state esaminate informazioni concernenti separatamente le variabili Altezza (X) e Peso (Y) per il data set STUDENTI. Di maggiore interesse può essere lo studio congiunto di tali variabili continue mediante l'analisi di una tabella doppia di frequenze nella quale, evidentemente, occorre raggruppare in classi di modalità i valori assunti da X e da Y (trattandosi di variabili continue).

Per semplicità, studieremo tali variabili solo per le $n = 1680$ studentesse, per le quali la variabile X varia tra 146 e 182 cm mentre la variabile Y varia tra 40 e 95 kg. Allora, suddividiamo l'intervallo $[146, 182]$ di X in 6 classi di eguale ampiezza e, similmente, suddividiamo l'intervallo $[40, 95]$ di Y in 5 classi di eguale ampiezza. Quindi, attribuiamo ciascuna studentessa alla casella specificata *contemporaneamente* sia dalle modalità di X che di Y, calcolando alla fine quante ricadono in ciascuna casella.

In tal modo, si perviene alla seguente distribuzione doppia delle frequenze assolute per la variabile doppia continua (X, Y):

$X//Y$	$40 \vdash - \dashv 51$	$51 - \dashv 62$	$62 - \dashv 73$	$73 - \dashv 84$	$84 - \dashv 95$	*Totali*
$146 \vdash - \dashv 152$	15	7	0	0	0	22
$152 - \dashv 158$	103	105	9	2	0	219
$158 - \dashv 164$	168	351	41	3	0	563
$164 - \dashv 170$	77	462	133	12	2	686
$170 - \dashv 176$	2	88	67	10	2	169
$176 - \dashv 182$	0	3	17	1	0	21
Totali	365	1016	267	28	4	1680

La tabella mostra che vi sono 15 studentesse la cui Altezza è compresa tra cm 146 (incluso) e cm 152 (incluso) le quali *contemporaneamente* hanno un Peso compreso tra kg 40 (incluso) e kg 51 (incluso). Similmente, vi sono 462 studentesse la cui Altezza è compresa tra cm 164 (più di cm 164) e cm 170 (incluso) le quali *contemporaneamente* hanno un Peso compreso tra kg 51 (più di kg 51) e kg 62 (incluso). Vi è, poi, una sola studentessa la cui Altezza è maggiore di cm 176 e inferiore o uguale a cm 182 la quale, *contemporaneamente*, ha un Peso compreso tra kg 73 (più di kg 73) e kg 84 (incluso).

Rispetto alle due distribuzioni semplici (dell'Altezza da sola, o del Peso da solo), la distribuzione doppia aggiunge delle nuove informazioni riguardanti la relazione tra le due variabili componenti. Come la tabella evidenzia, la maggior parte della distribuzione doppia è accentrata nella direzione della diagonale principale evidenziando che al crescere delle modalità di una delle due variabili cresce tendenzialmente anche l'altra, come era lecito attendersi dal legame tra Altezza e Peso. La nuvola dei punti conferma ulteriormente questa relazione.

Ovviamente, si può costruire (in perfetta analogia con il caso univariato discusso nell'esempio 3.3) un diagramma doppio a barre che rappresenta la traduzione grafica della precedente tabella a doppia entrata (Figura 7.2).

Si osservi che la conoscenza della distribuzione bivariata (cioè doppia) della variabile (X, Y) consente anche la conoscenza della distribuzione univariata (cioè semplice) delle due variabili componenti X e Y. Per esempio, dalla tabella doppia è agevole dedurre i totali di riga (per la X) ed i totali di colonna (per la Y), i quali sono le corrispondenti

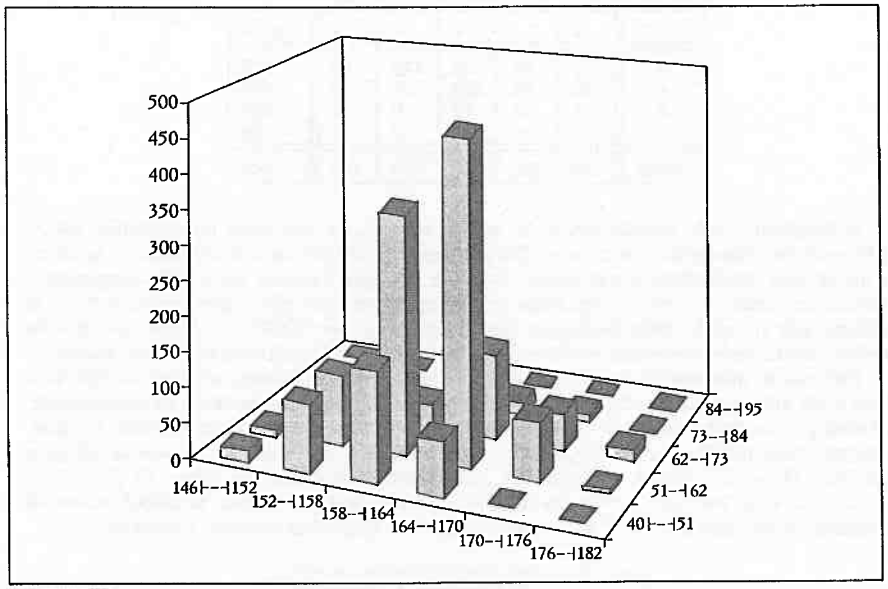

FIGURA 7.2. Diagramma a barre bivariato delle variabili «Peso» e «Altezza».

distribuzioni di frequenza delle variabili componenti, e sono riportate nelle due seguenti tabelle.

Altezza (X)	Frequenze assolute
$146 \vdash - \dashv 152$	22
$152 - \dashv 158$	219
$158 - \dashv 164$	563
$164 - \dashv 170$	686
$170 - \dashv 176$	169
$176 - \dashv 182$	21
Totale	1680

Peso (Y)	Frequenze assolute
$40 \vdash - \dashv 51$	365
$51 - \dashv 62$	1016
$62 - \dashv 73$	267
$73 - \dashv 84$	28
$84 - \dashv 95$	4
Totale	1680

Al contrario, conoscendo le distribuzioni di frequenza univariate delle due variabili componenti non è possibile dedurre l'unica distribuzione doppia da cui esse derivano. Per esempio, conoscere che 219 studentesse hanno un'Altezza compresa tra cm 152 e cm 158 ed inoltre che 365 studentesse hanno un Peso compreso tra kg 40 e kg 51 non consente di risalire al numero di studentesse che possiedono un'Altezza compresa tra cm 152 e cm 158 e *contemporaneamente* hanno un Peso compreso tra kg 40 e kg 51, come è invece possibile affermare esaminando la tabella della distribuzione doppia (dalla quale, subito, si deduce che vi sono 103 studentesse con tali caratteristiche).

□

Esempio 7.4

Dal data set STUDENTI, esaminiamo la variabile doppia mista (X, Y) così definita:
X = Numero di componenti della famiglia di appartenenza;
Y = Sesso.
La distribuzione doppia delle frequenze assolute è compendiata nella seguente tabella a doppia entrata, assieme ai totali per riga e per colonna che forniscono la distribuzione delle frequenze marginali assolute per la variabile quantitativa X e per la variabile qualitativa Y, rispettivamente.

Numero di Componenti	Sesso		Totali
	Maschi	Femmine	
1	12	2	14
2	26	29	55
3	142	190	332
4	422	635	1057
5	314	560	874
6	117	187	304
7	27	64	91
8	16	8	24
9	0	3	3
10	2	2	4
11	0	0	0
12	1	0	1
Totali	1079	1680	2759

Se ora, ordinatamente, si divide ciascuna delle frequenze assolute della precedente tabella per il numero totale degli studenti ($n = 2759$) si ottiene la tabella delle frequenze relative doppie:

Numero di Componenti	Sesso		Totali
	Maschi	Femmine	
1	0.00435	0.00072	0.00507
2	0.00942	0.01051	0.01993
3	0.05147	0.06887	0.12033
4	0.15295	0.23016	0.38311
5	0.11381	0.20297	0.31678
6	0.04241	0.06778	0.11018
7	0.00979	0.02320	0.03298
8	0.00580	0.00290	0.00870
9	0.00000	0.00109	0.00109
10	0.00072	0.00072	0.00145
11	0.00000	0.00000	0.00000
12	0.00036	0.00000	0.00036
Totali	0.39108	0.60892	1.00000

La tabella consente interessanti interpretazioni all'interno della popolazione esaminata e tra popolazioni differenti. Per esempio, il 15.3% degli studenti e il 23% delle studentesse provengono da nuclei familiari di 4 componenti; complessivamente le unità statistiche che vivono in nuclei familiari di 4 componenti costituiscono il 38.3% del totale, mentre poco più dello 0.1% degli studenti vive in famiglie di 9 componenti.

Ovviamente, esaminare la distribuzione del numero di componenti solo degli studenti oppure solo delle studentesse, in questo esempio, equivale a studiare la variabile condizionata $(X|Y = 0)$ oppure $(X|Y = 1)$, rispettivamente.

□

Gli esempi hanno mostrato come da una distribuzione doppia (X, Y) con k modalità distinte per la X ed h modalità distinte per la Y sia possibile derivare due distribuzioni univariate per le variabili componenti (le *distribuzioni marginali*), h *distribuzioni condizionate* della X per ciascuno dei valori della Y, k *distribuzioni condizionate* della Y per ciascuno dei valori della X. Questi studi hanno valore interpretativo di un certo rilievo come esemplificheremo, ora, in un caso di interesse economico.

Esempio 7.5

o È noto che sussiste una relazione tra il Reddito e il Consumo di una famiglia: la distribuzione di frequenza della variabile doppia (Reddito, Consumo), rilevata sull'unità statistica «famiglia», esplicita tale tipo di legame. Tuttavia, può essere interessante fissare un certo reddito familiare, per esempio 1000 Euro/mese, e verificare se ed in che misura variano i rispettivi consumi. In altri termini, si cerca di studiare la distribuzione della variabile Consumo delle famiglie limitatamente a quelle per le quali la variabile Reddito è pari a 1000 Euro/mese. Un discorso analogo può essere fatto dopo aver fissato la variabile Reddito a 2000 Euro/mese, a 2500 Euro/mese, e così via.

L'analisi economica e l'esperienza di ciascuno mostrano che certamente il Consumo varia con il Reddito, che a Redditi più elevati corrispondono tendenzialmente Consumi più elevati, ma anche che la variabilità dei Consumi si modifica al crescere del livello dei Redditi. Infatti, quando il livello del Reddito è relativamente basso, i Consumi sono abbastanza omogenei nella popolazione dovendo soddisfare anzitutto le esigenze primarie della famiglia (cibo, casa, sanità, etc.); per contro, a livelli più elevati della variabile Reddito, emergono le scelte e le preferenze individuali (cultura, tempo libero, sport, beni voluttuari, etc.) che generano maggiore variabilità nella distribuzione dei Consumi.

Queste considerazioni fanno emergere due fatti: da un lato, il Consumo medio varia positivamente con il Reddito; dall'altro la variabilità del Consumo non è costante al variare del Reddito ma cresce con esso.

□

7.3 Distribuzioni condizionate e indipendenza

Formalizziamo, in questo paragrafo, alcuni dei concetti già esemplificati in precedenza, limitandoci sempre al caso bivariato. La generalizzazione al caso multivariato è agevole dal punto di vista concettuale, ma la notazione e la rappresentazione si complicano notevolmente, per cui continueremo a discutere delle sole variabili doppie.

Data una distribuzione doppia (X, Y), se fissiamo il valore x_i per la variabile X ed esaminiamo la distribuzione di Y limitatamente alle unità statistiche che possiedono quel valore x_i per la variabile X, otteniamo la distribuzione condizionata di Y dato x_i cioè la *variabile condizionata* $(Y|X = x_i)$. Quando non vi è motivo di confusione, tale distribuzione si indica talvolta anche con $Y|x_i$.

In modo perfettamente analogo, se fissiamo un valore y_j per la variabile Y ed esaminiamo la distribuzione di X limitatamente alle unità statistiche che possiedono quel valore y_j per la variabile Y, otteniamo la distribuzione condizionata di X dato y_j cioè la *variabile condizionata* $(X|Y = y_j)$, indicata anche con $X|y_j$.

Ora, tutte le variabili condizionate sono variabili univariate, per cui esse so-

no note se conosciamo i valori che assumono e le rispettive frequenze. Pertanto, per un fissato x_i la distribuzione di frequenza della variabile condizionata $Y|(X = x_i)$ e per un fissato y_j la distribuzione di frequenza della variabile condizionata $X|(Y = y_j)$ saranno così specificate:

Modalità di $Y\|(X = x_i)$	Frequenze assolute	Frequenze relative
y_1	n_{i1}	$n_{i1}/n_{i.}$
y_2	n_{i2}	$n_{i2}/n_{i.}$
...	...	...
y_j	n_{ij}	$n_{ij}/n_{i.}$
...	...	...
y_h	n_{ih}	$n_{ih}/n_{i.}$
Totali	$n_{i.}$	1

Modalità di $X\|(Y = y_j)$	Frequenze assolute	Frequenze relative
x_1	n_{1j}	$n_{1j}/n_{.j}$
x_2	n_{2j}	$n_{2j}/n_{.j}$
...	...	...
x_i	n_{ij}	$n_{ij}/n_{.j}$
...	...	...
x_k	n_{kj}	$n_{kj}/n_{.j}$
Totali	$n_{.j}$	1

Va sottolineato, tuttavia, come la conoscenza delle distribuzioni univariate non è sufficiente a ricostruire la distribuzione doppia (tranne in casi molto particolari, come vedremo). Infatti, ciò che è veramente nuovo in una distribuzione multivariata è l'informazione connessa al *contemporaneo* presentarsi delle modalità di una variabile con quelle di un'altra variabile. Tale aspetto si manifesta nel disporsi delle frequenze in una tabella multipla in modo che certe coppie di modalità si presentino più frequentemente di altre. Assume, pertanto, notevole interesse studiare questo «disporsi» delle frequenze doppie secondo certi addensamenti che manifestano evidentemente una più elevata predisposizione di certe coppie di modalità di (X, Y) di presentarsi assieme, e quindi un certo legame.

Per evidenziare queste relazioni si possono introdurre tipi di indicatori che utilizzano le sole frequenze oppure indicatori che includono nella sintesi anche le modalità delle variabili (nel caso che queste siano quantitative). Nel primo caso si parla di *misure di connessione*, e saranno esaminate nel prossimo paragrafo, mentre nel secondo si parla di *misure di correlazione*, e saranno studiate nel paragrafo 7.5.

Gli indici di connessione sono prevalentemente (ma non esclusivamente) applicati allo studio di variabili qualitative mentre gli indici di correlazione possono essere calcolati solo tra due variabili entrambe quantitative (perché utilizzano esplicitamente le modalità).

Ora, per studiare le relazioni tra le componenti di una generica variabile doppia (X, Y) ci si può chiedere qual è la situazione estrema di assenza di qualsiasi legame tra tali variabili. In effetti, tra X e Y vi è assenza di qualsiasi legame se vi è *indipendenza*, cioè se qualsiasi valore assunto da X non modifica la distribuzione di frequenza della variabile Y (e viceversa, qualunque valore assunto da Y non modifica la distribuzione di frequenza della variabile X).

Allora, *l'indipendenza implica che la conoscenza di un valore di una delle due variabili non deve aggiungere alcuna informazione sulla distribuzione dell'altra.*

In modo formale, questo equivale a dire che ***le componenti X e Y di una variabile doppia (X, Y) sono indipendenti se e solo se le distribuzioni delle frequenze relative condizionate di X e di Y sono costanti.***

Quindi, dovrà verificarsi per ogni valore x_i di X che:

$$\frac{n_{i1}}{n_{.1}} = \frac{n_{i2}}{n_{.2}} = \cdots = \frac{n_{ih}}{n_{.h}} = \frac{n_{i.}}{n}$$

e, similmente, dovrà verificarsi per ogni valore y_j di Y che:

$$\frac{n_{1j}}{n_{1.}} = \frac{n_{2j}}{n_{2.}} = \cdots = \frac{n_{kj}}{n_{k.}} = \frac{n_{.j}}{n}\,.$$

Se ne deduce agevolmente che questo accade quando, e solo quando, per ogni coppia di modalità (x_i, y_j), tutte le frequenze doppie e marginali sono legate da una proporzionalità per righe e per colonne.

Pertanto, se le componenti X e Y di una variabile doppia (X, Y) sono indipendenti allora sussistono le seguenti relazioni:

$$n_{ij} = \frac{n_{i.}\, n_{.j}}{n}, \qquad \forall (i,j),\ i = 1, 2, \ldots, k;\ j = 1, 2, \ldots, h.$$

Per convincersi della necessità di tali relazioni, esaminiamo in dettaglio un caso ipotetico di indipendenza tra due variabili qualitative, come nell'esempio seguente.

Esempio 7.6

Supponiamo che su una popolazione di $n = 200$ adulti si siano raccolte informazioni sulla variabile doppia (X, Y), dove:

X = Sesso (nelle modalità: Maschi, Femmine);
Y = Bibita preferita (nelle modalità: A, B, C).

Si è, quindi, ottenuta la seguente distribuzione doppia di frequenze assolute.

$\downarrow X \ // \ Y \rightarrow$	A	B	C	Totali
Maschi	40	28	12	80
Femmine	60	42	18	120
Totali	100	70	30	200

Si nota subito che, esaminando il problema rispetto alla variabile X, i maschi e le femmine rappresentano il 40% e il 60%, rispettivamente, della popolazione e che le proporzioni di preferenza per ciascuna delle tre bibite sono esattamente ed ordinatamente sempre il 40% e il 60%.

Da un altro punto di vista, esaminando il problema rispetto alla variabile Y, il gradimento per le tre bibite risulta nelle percentuali del 50% per A, del 35% per B e del 15% per C sull'intera popolazione e tali proporzioni sono ugualmente rispettate sia per i maschi che per le femmine, come è agevole constatare.

Quindi, la distribuzione della preferenza delle bibite condizionata al sesso è costante (al variare del sesso) e la distribuzione del sesso condizionata alla bibita preferita è costante (al

variare della bibita). Questo caso (teorico) costituisce un esempio di perfetta indipendenza tra le due variabili: infatti, l'informazione che si è maschio (femmina) non modifica in alcun modo quella sulla bibita preferita e, similmente, l'informazione sulla bibita preferita non aggiunge alcuna informazione sul sesso della unità statistica.

Come è agevole constatare, tutte le frequenze doppie rispettano la proporzione rispetto ai totali di riga e di colonna e, di conseguenza, la relazione tra frequenze doppie e marginali. Infatti, si ha:

$$40 = \frac{100 \times 80}{200};\ 28 = \frac{70 \times 80}{200};\ 12 = \frac{30 \times 80}{200};\ldots;\ 18 = \frac{30 \times 120}{200}\ .$$

□

Quindi, la situazione estrema di nessun legame tra le variabili è quella di indipendenza, per cui qualsiasi distribuzione differente da quella di indipendenza implica un qualche legame, per quanto debole questo legame possa essere. È logico allora misurare il legame tra le variabili mediante una sorta di distanza tra le frequenze doppie realmente *osservate* nella popolazione in esame e le frequenze doppie *ipotetiche* che si sarebbero dovute osservare nel caso vi fosse stata una perfetta indipendenza tra le variabili. Maggiore è tale distanza, maggiore sarà il legame tra le variabili considerate.

7.4 Una misura di connessione tra variabili doppie

Data una variabile doppia (X, Y) di cui conosciamo le frequenze assolute doppie n_{ij}, per $i = 1, 2, \ldots, k; j = 1, 2, \ldots, h$. Indichiamo con:

$$c_{ij} = \frac{n_{i.}\ n_{.j}}{n}, \quad i = 1, 2, \ldots, k;\ j = 1, 2, \ldots, h,$$

le corrispondenti *frequenze assolute attese* calcolate sotto l'assunzione di una perfetta indipendenza tra X e Y. Tenuto conto di quanto si è discusso nel paragrafo precedente, una misura di connessione crescerà con la distanza tra n_{ij} e c_{ij}.

Nel 1900, Karl Pearson propose il seguente ***indice quadratico di connessione***:

$$X^2 = \sum_{i=1}^{k}\sum_{j=1}^{h} \frac{(n_{ij} - c_{ij})^2}{c_{ij}} = n\left(\sum_{i=1}^{k}\sum_{j=1}^{h} \frac{n_{ij}^2}{n_{i.}\ n_{.j}} - 1\right).$$

La seconda espressione è solo uno sviluppo algebrico della prima, ed è molto utile in fase di programmazione. Infatti, tale espressione (essendo esclusivamente basata sulle frequenze osservate, doppie e marginali) non richiede il calcolo preliminare delle frequenze c_{ij}, attese sotto l'ipotesi di indipendenza.

Poiché si dimostra che: $X^2 \leq n\{\min(k-1, h-1)\}$, è stato proposto l'***indice normalizzato di Cramér:***

$$\nu = \sqrt{\frac{X^2}{n\{\min(k-1, h-1)\}}}\ .$$

per il quale si verifica sempre: $0 \leq \nu \leq 1$.

Quando sia la X che la Y possiedono due sole modalità si parla di *tabella tetracorica* (perché le sole frequenze essenziali sono 4).

o La situazione è piuttosto diffusa nelle analisi reali, soprattutto laddove si cerca una connessione tra due variabili qualitative; per esempio, tra trattamento e guarigione; tra comparsa di una malattia e tipologia della dieta; tra frequenza ai corsi e superamento dell'esame; tra livelli di mortalità degli incidenti stradali ed uso delle cinture di sicurezza; tra abilità manuali e superamento dei test attitudinali, etc.

Per una tabella tetracorica, è possibile dimostrare – con ovvia simbologia per le frequenze assolute – che gli indici di Pearson e di Cramér si semplificano, rispettivamente, nel modo seguente:

$$X^2 = \frac{n\,(n_{11}n_{22} - n_{12}n_{21})^2}{n_{.1}n_{.2}n_{1.}n_{2.}}; \qquad \nu = \frac{|n_{11}n_{22} - n_{12}n_{21}|}{\sqrt{n_{.1}n_{.2}n_{1.}n_{2.}}}.$$

Esempio 7.7

Supponiamo di voler misurare l'eventuale connessione tra l'essere favorevoli ad una certa proposta di legge ed il sesso dell'intervistato mediante lo studio delle risposte fornite da una popolazione di $n = 303$ unità statistiche.

La tabella seguente mostra le frequenze osservate e (tra parentesi) le corrispondenti frequenze attese, calcolate sotto l'ipotesi teorica che, a parità di totali marginali, vi sia perfetta indipendenza tra le variabili.

Risposta // Sesso	Maschi	Femmine	Totali
Favorevoli	38(30.68)	74(81.32)	112
Non-favorevoli	45(52.32)	146(138.68)	191
Totali	83	220	303

Per esempio, sono stati osservati 38 maschi favorevoli mentre, se vi fosse stata perfetta indipendenza tra le variabili, occorreva osservarne circa 31 (=30.68), e così via. Come la tabella evidenzia, vi è una certa dissimilarità tra le frequenze osservate e quelle che ci si poteva attendere nel caso di indipendenza.

Una misura adeguata di tale discrepanza è l'indice X^2 il cui calcolo è immediato:

$$X^2 = \frac{(38 - 30.68)^2}{30.68} + \frac{(74 - 81.32)^2}{81.32} + \frac{(45 - 52.32)^2}{52.32} + \frac{(146 - 138.68)^2}{138.68} = 3.8159$$

Trattandosi di una tabella tetracorica si può anche utilizzare il calcolo abbreviato mediante la formula:

$$X^2 = \frac{303(38 \times 146 - 45 \times 74)^2}{83 \times 220 \times 112 \times 191} = 3.8159$$

Infine, l'indice normalizzato di Cramér è pari a:

$$\nu = \frac{|38 \times 146 - 45 \times 74|}{\sqrt{83 \times 220 \times 112 \times 191}} = 0.11221$$

Tali risultati mostrano che la connessione tra le due variabili esaminate esiste (perché $X^2 > 0$) ma è comunque di debole intensità (perché l'indice ν è piuttosto distante dal suo massimo, che è 1).

□

7.5 Correlazione tra variabili statistiche

Quando le componenti della variabile doppia (X, Y) sono entrambe variabili quantitative, allora è preferibile utilizzare una misura esplicita del loro legame che coinvolge anche le modalità. In effetti, come vedremo, il coefficiente che introdurremo in questo paragrafo misura solo un particolare legame statistico, quello di tipo lineare, cioè esprimibile mediante una relazione di primo grado tra le variabili X e Y. Tuttavia, il suo studio è preliminare ed importante rispetto a qualsiasi altro tipo di eventuale legame tra variabili quantitative.

o La discussione che stiamo sviluppando mira alla costruzione di uno specifico indice statistico; tuttavia, essa deriva in modo «naturale» dal ragionamento comune e dalle deduzioni razionali delle scienze fisiche, economiche, sociali, etc. In tutti i settori, infatti, l'investigazione comincia con la rappresentazione grafica dei risultati ottenuti dalla rilevazione di due variabili per le stesse unità statistiche, allo scopo di «avere un'idea» circa l'eventuale legame che potrebbe sussistere tra tali variabili: ciò ha una funzione descrittiva, inferenziale, interpretativa e previsiva, come evidenziano gli esempi seguenti:

- Registrando la velocità X e lo spazio di frenata Y di n autoveicoli sottoposti a collaudo, ci si attende che al crescere della velocità sarà necessario maggiore spazio per la frenata, anche se nessuno potrà credere che sussista una perfetta relazione matematica tra X e Y.
- Nella rilevazione dei redditi X e consumi Y di n famiglie ci si attende una relazione crescente tra X e Y perché l'esperienza comune e la teoria economica concordano sul fatto che, entro certi limiti e con necessari margini di variabilità, i consumi familiari crescono in funzione del reddito disponibile.
- Tra il numero di anni X trascorsi all'Università e il voto finale di laurea Y sussiste una relazione inversa perché al crescere del tempo impiegato all'Università peggiora la qualità e la produttività nello studio per una serie di motivi (stanchezza psicologica, difficoltà nella frequenza dei corsi, necessità di lavorare, etc.) che concorrono a determinare un voto finale di laurea comparativamente più basso.

Supponiamo di possedere una serie doppia (x_i, y_i), per $i = 1, 2, \ldots, n$, ottenuta in uno studio sulla variabile doppia (X, Y). La rappresentazione tradizionale sugli assi cartesiani, nella quale a ciascun soggetto si fa corrispondere un punto che ha per ascissa x_i e per ordinata y_i (cioè la «nuvola di punti» già utilizzata in precedenza), è un primo modo per verificare se, ed in quale misura, sussista una relazione tra le variazioni di X e quelle di Y.

Poiché le variabili componenti X ed Y della variabile doppia avranno presumibilmente posizione e variabilità differenti, è conveniente riferirsi alle variabili standardizzate:

$$Z_x = \frac{X - \mu_x}{\sigma_x}; \qquad Z_y = \frac{Y - \mu_y}{\sigma_y};$$

che daranno luogo, nella popolazione, alle seguenti coppie di osservazioni standardizzate:

$$z_{x,i} = \frac{x_i - \mu_x}{\sigma_x}; \qquad z_{y,i} = \frac{y_i - \mu_y}{\sigma_y}; \quad i = 1, 2, \ldots, n.$$

Le variabili Z_x e Z_y hanno media 0 e varianza 1 (vedi paragrafo 6.2), per cui sono rappresentabili come una nuvola di punti disposta all'origine degli assi con variabilità uguale ad 1 per entrambe le variabili. Quindi, il grafico delle due variabili componenti standardizzate (cioè il grafico di $z_{y,i}$ rispetto a $z_{x,i}, i = 1, 2, \ldots, n$) è certamente utile per visualizzare se esiste una relazione tra le modalità di X e quelle di Y.

Le variabili standardizzate assumono valori positivi e negativi; pertanto, vi sarà concordanza tra X e Y tutte le volte che a valori positivi di Z_x corrispondono per lo più valori positivi di Z_y e a valori negativi di Z_x corrispondono per lo più valori negativi di Z_y. Ciò vuol dire che a valori di X superiori alla propria media (e quindi positivi di Z_x) corrispondono tendenzialmente valori di Y superiori alla propria media (e quindi positivi per Z_y), mentre a valori di X inferiori alla propria media (e quindi negativi di Z_x) corrispondono tendenzialmente valori di Y inferiori alla propria media (e quindi negativi per Z_y). Pertanto, nel caso di concordanze tra X e Y, i prodotti $(Z_x\, Z_y)$ saranno, per la gran parte delle osservazioni, di segno positivo perché costituiti da fattori «positivi×positivi» e «negativi×negativi».

Per contro, vi sarà discordanza tra X e Y tutte le volte che a valori positivi di Z_x corrisponderanno per lo più valori negativi di Z_y e a valori negativi di Z_x corrisponderanno per lo più valori positivi di Z_y. Ciò vuol dire che a valori di X inferiori alla propria media (e quindi negativi di Z_x) corrispondono tendenzialmente valori di Y superiori alla propria media (e quindi positivi per Z_y), mentre a valori di X superiori alla propria media (e quindi positivi di Z_x) corrispondono tendenzialmente valori di Y inferiori alla propria media (e quindi negativi per Z_y). Pertanto, nel caso di discordanza tra X e Y, i prodotti $(Z_x\, Z_y)$ saranno, per la gran parte delle osservazioni, di segno negativo perché costituiti principalmente da fattori «positivi×negativi» e «negativi×positivi».

Una misura sintetica che tenga conto complessivamente di tali quantità è la media aritmetica di tali prodotti: $z_{x,i}\, z_{y,i}, i = 1, 2, \ldots, n$. Infatti, tale sintesi riassume tendenzialmente il complesso delle variazioni congiunte di X e Y, al netto degli effetti causati dalla loro posizione e variabilità.

Si introduce così il ***coefficiente di correlazione lineare di Bravais-Pearson***, indicato nella letteratura con vari simboli: $\rho, \rho_{xy}, \rho(X,Y), Corr(X,Y)$, e definito da una delle seguenti espressioni alternative:

$$\rho = \frac{1}{n}\sum_{i=1}^{n}(z_{x,i}\ z_{y,i}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\mu_x}{\sigma_x}\right)\left(\frac{y_i-\mu_y}{\sigma_y}\right),$$

ovvero:

$$\rho = Corr(X,Y) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{\sigma_x\,\sigma_y} = \frac{\sigma_{xy}}{\sigma_x\sigma_y}.$$

La quantità al numeratore – indicata con σ_{xy} – è la media aritmetica del prodotto degli scarti di X e di Y e si chiama ***covarianza***, perché è una misura della contemporanea variazione di X e Y in rapporto alle rispettive medie.

Poiché il denominatore di ρ è positivo, la covarianza ha lo stesso segno del coefficiente di correlazione. Tuttavia, la covarianza dipende dalle unità di misura sia di X che di Y per cui non è utile a fini di confronto; invece, il coefficiente di correlazione (essendo una sintesi di quantità standardizzate) non presenta tale caratteristica e, quindi, si presta maggiormente a misurare il legame lineare tra variabili quantitative, che possono essere espresse in qualsiasi unità di misura.

L'importanza del coefficiente di correlazione lineare ρ deriva dalle numerose proprietà statistiche di cui gode e tra le quali ci limitiamo a segnalare (senza dimostrarle) quelle più rilevanti ai fini del nostro percorso, aggiungendo poi qualche commento per illustrarne il significato operativo.

1. $Corr(X,Y) = Corr(Y,X)$.
2. $Corr(X,X) = 1$.
3. $-1 \le Corr(X,Y) \le +1$.
4. $Corr(aX+b, cY+d) = segno(a\times c)\,Corr(X,Y)$.
5. $Corr(X,Y) = \pm 1 \iff Y = \beta_0 \pm \beta_1 X$.
6. X e Y *indipendenti* $\Longrightarrow Corr(X,Y) = 0$.

L'insieme di tali proprietà mostra che il coefficiente di correlazione lineare è una misura della forza del legame lineare (=di primo grado) che esiste tra due variabili quantitative. Esso cresce, in valore assoluto, quanto più la nuvola dei punti si addensa attorno ad una ipotetica linea retta che l'attraversa.

A fini interpretativi, la sua importanza deriva dal fatto che esso è invariante rispetto a trasformazioni di scala e di livello, è simmetrico ed è normalizzato in $[-1,+1]$: quindi, è immediato interpretare il suo valore numerico per misurare il legame lineare tra variabili quantitative.

Commentiamo, brevemente, il significato statistico di tali proprietà:

- La prima proprietà esprime la simmetria del coefficiente ρ.
- La seconda proprietà indica che la correlazione tra una variabile X e sé stessa è massima perché l'espressione $Y = X$ è perfettamente lineare.
- La terza proprietà esplicita il fatto che ρ è sempre compreso tra -1 e $+1$, per cui è un *indice normalizzato*.
- La quarta proprietà mostra che se le variabili X e Y sono sottoposte ad una trasformazione di primo grado (per qualsiasi costanti reali a, b, c, d) il coefficiente di correlazione lineare non cambia, a meno che il segno di $a \times c$ non sia negativo: in tal caso, si modifica solo il segno di ρ.
- La quinta proprietà esplicita il fatto che il coefficiente ρ raggiunge i suoi valori estremi, cioè ± 1, se e solo se tra le variabili X e Y esiste una perfetta relazione lineare esprimibile, quindi, tramite una equazione di primo grado, cioè: $Y = \beta_0 \pm \beta_1 X$. Ciò implica che, quando $\rho = \pm 1$, allora per ciascuna unità statistica sussiste una perfetta relazione lineare del tipo: $y_i = \beta_0 + \beta_1 x_i, i = 1, 2, \ldots, n$. In tal caso, quindi, nota la relazione lineare, dalla conoscenza di X si può risalire perfettamente alla conoscenza di Y, e viceversa. Ovviamente, tale relazione deterministica è inverosimile per fenomeni reali ma viene qui considerata come caso estremo di prevedibilità perfetta del valore di una variabile Y tramite quello di un'altra variabile X mediante una relazione di primo grado (il Capitolo 18 approfondirà questi aspetti).
- La sesta proprietà specifica un aspetto importante della correlazione perché mostra che se tra due variabili non vi è *nessun tipo di relazione* – cioè se X e Y sono indipendenti – allora, a maggior ragione, non vi può essere una relazione di primo grado (e quindi la correlazione è 0, e in tal caso si dice che X e Y sono *variabili incorrelate*). Ovviamente, non vale il viceversa, perché se due variabili sono incorrelate vuol dire solo che non vi è un legame di primo grado, ma ve ne possono essere di altro tipo. In altri termini, ciò si esprime dicendo che *l'indipendenza implica la incorrelazione ma non viceversa.*

Questi risultati mostrano che ρ è un indice di correlazione lineare perché misura il legame lineare tra Y e X. La massima correlazione negativa si ha quando $\rho = -1$, il che implica che tutte le coppie di osservazioni giacciono su una retta con pendenza negativa; invece, la massima correlazione positiva si ha per $\rho = +1$, il che implica che tutte le coppie di osservazioni giacciono su una retta con pendenza positiva.

Pertanto, quando $|\ \rho\ |$ raggiunge il suo massimo valore di 1 si ha una perfetta relazione lineare tra X e Y, ed in effetti, in tali casi, si tratta di punti perfettamente allineati.

I due esempi seguenti mostrano sia per alcuni casi simulati che per alcuni casi reali, il valore numerico che assume il coefficiente ρ.

- Quando i dati sono disponibili in serie: $(x_i, y_i), i = 1, 2, \ldots, n$, e non si possiede un adeguato software per ottenere ρ, basta calcolare esclusivamente le seguenti somme:

$$S_{1x} = \sum x_i;\ S_{1y} = \sum y_i;\ S_{2x} = \sum x_i^2;\ S_{2y} = \sum y_i^2;\ S_{xy} = \sum x_i y_i.$$

Infatti, si dimostra che il coefficiente di correlazione lineare è uguale a:

$$\rho = \frac{n\,S_{xy} - S_{1x}\,S_{1y}}{\sqrt{(n\,S_{2x} - (S_{1x})^2)\,(n\,S_{2y} - (S_{1y})^2)}}.$$

- Quando i dati sono disponibili in seriazione, cioè mediante una tabella a doppia entrata ove a ciascuna coppia (x_i, y_j) corrisponde la frequenza assoluta n_{ij}, allora il calcolo di ρ avviene mediante la seguente formulazione:

$$\rho = \frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{k}\sum_{j=1}^{h}(x_i - \mu_x)(y_j - \mu_y)\,n_{ij}}{\sqrt{\left\{\dfrac{1}{n}\displaystyle\sum_{i=1}^{k}(x_i - \mu_x)^2\,n_{i\cdot}\right\}\left\{\dfrac{1}{n}\displaystyle\sum_{j=1}^{h}(y_j - \mu_y)^2\,n_{\cdot j}\right\}}}.$$

In tale espressione, al numeratore vi è la covarianza per variabili (X, Y) organizzate in una tabella a doppia entrata e, similmente, al denominatore sono riportate (nelle due parentesi sotto radice) le espressioni per la varianza di X e Y, rispettivamente.

Esempio 7.8

Esplicitiamo, in questo esempio, i calcoli necessari per la determinazione di ρ nel caso si disponga di dati sulle variabili (X, Y) mediante una serie di $n = 25$ misurazioni sulle variabili X = Altezza (in cm) e Y = Peso (in kg) ottenute su un gruppo di adulti. La tabella esplicita i calcoli necessari (agevolmente derivabili mediante un foglio elettronico Excel, per esempio).

i	x_i	y_i	x_i^2	y_i^2	$x_i\,y_i$
1	184	82	33856	6724	15088
2	180	90	32400	8100	16200
3	168	58	28224	3364	9744
4	170	57	28900	3249	9690
5	169	57	28561	3249	9633
6	170	63	28900	3969	10710
7	178	83	31684	6889	14774
8	170	78	28900	6084	13260
9	173	65	29929	4225	11245
10	160	61	25600	3721	9760
11	167	51	27889	2601	8517
12	187	77	34969	5929	14399
13	165	67	27225	4489	11055
14	185	69	34225	4761	12765
15	161	67	25921	4489	10787
16	155	45	24025	2025	6975
17	156	52	24336	2704	8112
18	157	47	24649	2209	7379
19	165	55	27225	3025	9075
20	180	70	32400	4900	12600
21	165	54	27225	2916	8910
22	170	60	28900	3600	10200
23	167	58	27889	3364	9686
24	175	74	30625	5476	12950
25	160	60	25600	3600	9600
Totali	$S_{1x} = 4237$	$S_{1y} = 1600$	$S_{2x} = 720057$	$S_{2y} = 105662$	$S_{xy} = 273114$

Sfruttando la precedente espressione per il calcolo di ρ quando i dati sono disponibili in serie, e le somme ottenute dalla precedente tabella, si ha:

$$\rho = \frac{25 \times 273114 - 1600 \times 4237}{\sqrt{(25 \times 105662 - (1600)^2)\,(25 \times 720057 - (4237)^2)}} = 0.76761.$$

I dati evidenziano, quindi, una consistente positiva correlazione tra Peso e Altezza nel gruppo considerato. □

Esempio 7.9

Riprendiamo, ora, i dati della seriazione sulla variabile (X, Y) contenuti nella tabella a doppia entrata di cui all'esempio 7.1 e che qui ripetiamo per comodità espositiva. Essi riguardano la distribuzione doppia di frequenza del numero di auto (variabile X) e del numero di componenti (variabile Y) di una famiglia, ottenuta da una popolazione di $n = 1000$ unità statistiche.

X//Y	1	2	3	4	5	*Totali*
0	10	20	20	150	50	250
1	85	85	330	50	50	600
2	5	85	10	0	0	100
3	0	10	40	0	0	50
Totali	100	200	400	200	100	1000

Le somme necessarie per il calcolo di ρ sono le seguenti:

$$S_{1x} = \sum_{i=1}^{k} x_i\, n_{i.} = 0 \times 250 + 1 \times 600 + 2 \times 100 + 3 \times 50 = 950;$$

$$S_{1y} = \sum_{j=1}^{h} y_j\, n_{.j} = 1 \times 100 + 2 \times 200 + 3 \times 400 + 4 \times 200 + 5 \times 100 = 3000;$$

$$S_{2x} = \sum_{i=1}^{k} x_i^2\, n_{i.} = 0^2 \times 250 + 1^2 \times 600 + 2^2 \times 100 + 3^2 \times 50 = 1450;$$

$$S_{2y} = \sum_{j=1}^{h} y_j^2\, n_{.j} = 1^2 \times 100 + 2^2 \times 200 + 3^2 \times 400 + 4^2 \times 200 + 5^2 \times 100 = 10200;$$

$$S_{xy} = \sum_{i=1}^{k} \sum_{j=1}^{h} x_i\, y_j\, n_{ij} = (0 \times 1) \times 10 + (1 \times 1) \times 85 + \cdots + (3 \times 5) \times 0 = 2525.$$

Pertanto, applicando la formula semplificatrice riportata nel riquadro del testo, si ha:

$$\rho = \frac{1000 \times 2525 - 950 \times 3000}{\sqrt{(1000 \times 1450 - (950)^2)\,(1000 \times 10200 - (3000)^2)}} = -0.40096.$$

Come si vede, nella popolazione considerata, si registra un sensibile legame negativo tra numero di componenti e auto possedute dalla famiglia, cioè al crescere del numero dei componenti della famiglia vi è una tendenza alla decrescita del numero di auto possedute. □

Esempio 7.10

Per interpretare il valore di ρ nei casi concreti, presentiamo nella Figura 7.3 alcuni gruppi di osservazioni ove è esplicitato il valore del coefficiente di correlazione lineare. Essi mostrano come nel passaggio da un valore negativo di ρ a valori positivi la pendenza si inverta e come l'addensamento dei punti è tanto più vicino ad una retta che li attraversa quanto più il valore assoluto di ρ è elevato. Quando, invece, il coefficiente di correlazione ρ è attorno allo 0 si osserva una nuvola dei punti senza una direzione definita: infatti, in tal caso, mediamente la variabile Y non varia al variare della variabile X, come ci si attende in assenza di un legame lineare. □

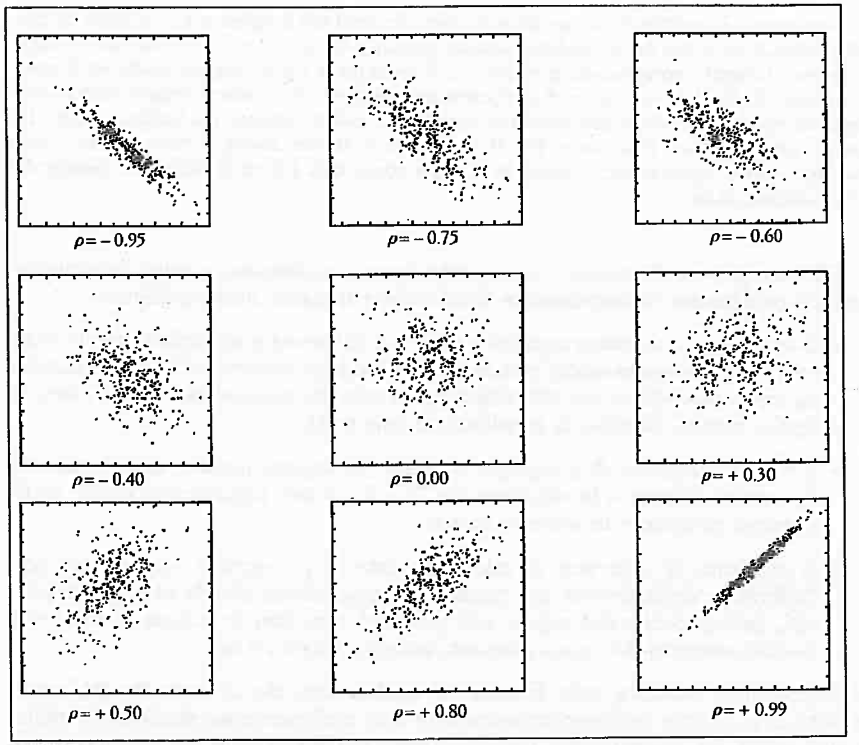

FIGURA 7.3. Esemplificazioni di variabili doppie e relativo coefficiente di correlazione lineare.

Esempio 7.11

Nella tabella seguente, presentiamo alcuni valori di ρ in corrispondenza di sottoinsiemi dei data set e delle variabili appresso indicate.

Data set	*Variabile X*	*Variabile Y*	*Corr(X,Y)*
STUDENTI	Peso dei Maschi	Altezza dei Maschi	+0.59
STUDENTI	Peso delle Femmine	Altezza delle Femmine	+0.49
STUDENTI	Età dei Maschi	Altezza dei Maschi	−0.10
STUDENTI	Età delle Femmine	Altezza delle Femmine	−0.11
CAMELLIA	Lunghezza della foglia	Larghezza della foglia	+0.52
CAMELLIA	Peso vivo	Peso secco	+0.91

Mentre sono ovvi i valori positivi di ρ per caratteri antropometrici e per caratteri fisici (essi, per ogni persona, crescono nella stessa direzione, con una variabilità individuale sensibile, sicché pur essendo notevoli sono abbastanza lontani da +1), la correlazione negativa tra Età ed Altezza degli studenti si spiega con il fatto che universitari di età più elevata provengono da generazioni precedenti la cui altezza è mediamente più bassa.

Similmente, la correlazione (positiva ma non elevata) tra lunghezza e larghezza di una foglia indica il fatto che foglie tendenzialmente grandi/piccole lo sono nelle due dimensioni.

Invece, l'elevata correlazione positiva tra il peso della foglia appena colta ed il peso della stessa foglia dopo un mese di essiccazione indica che la quota di evaporazione dell'acqua in essa contenuta è praticamente costante e molto regolare da foglia a foglia. In breve, è come se tra il Peso secco Y e il Peso vivo X di una foglia vi fosse «quasi» una relazione lineare, come schematizzato in R 7.2 e come sarà poi confermato nel paragrafo 18.8 (Esempio 18.6).

□

Gli esempi e la discussione precedente hanno evidenziato alcune interpretazioni del coefficiente di correlazione lineare che riteniamo utile riassumere.

- Il *segno di* ρ può essere negativo, nullo o positivo ed è immediatamente connesso alla direzione media complessiva del legame lineare così come misurato da una retta ideale che attraversa la nuvola dei punti delle osservazioni; il segno, quindi, esprime la pendenza di tale retta.
- Il *valore numerico di* ρ esprime la forza del legame lineare, cioè la misura di quanto intensa è la relazione tra X e Y, e tale legame può essere forte in senso positivo o in senso negativo.
- Il *quadrato di* ρ (ovvero il valore assoluto di ρ) esprime una misura dell'aderenza della nuvola dei punti alla retta ideale che la attraversa perché, prescindendo dal segno, tali quantità crescono in misura diretta con l'addensamento dei valori osservati attorno a tale retta.

Le osservazioni simulate nella Figura 7.3 confermano che al crescere del valore assoluto di ρ tutte le osservazioni assumono una configurazione sempre più vicina ad una retta (la cui direzione è determinata dal segno e la cui inclinazione è determinata dal valore numerico di ρ).

Da queste considerazioni deriva anche un altro importante significato per il coefficiente di correlazione lineare che, infatti, può essere interpretato come *grado di prevedibilità lineare* di una variabile tramite un'altra. Infatti, se il valore assoluto di ρ è elevato allora la conoscenza di X «aiuta» nella possibile determinazione di Y perché esso si situerà, con bassa variabilità, attorno al valore determinato dalla retta che attraversa la nuvola dei punti (si osservino le situazioni evidenziate nella Figura 7.3). Invece, quanto più $\rho \to 0$ tanto più i valori di Y saranno variabili attorno alla retta parallela all'asse X: in quest'ultima situazione, evidentemente, la conoscenza di un valore per X non aiuta a conoscere il corrispondente valore per Y, che sarà quasi interamente determinato dalla sua media costante e da una intrinseca variabilità non spiegabile (in tal caso) da X.

o Accanto a questi aspetti, ci sembra importante sottolineare alcune cautele da tenere presenti nella interpretazione di ρ, soprattutto quando il calcolo viene ottenuto tramite elaborazioni automatiche, che potrebbero essere ottenute senza nemmeno osservare i dati e riflettere sulla loro natura. Tali perplessità riguardano i problemi generati da alcune situazioni che ora, brevemente, richiamiamo.

- ***Correlazione a blocchi.*** Il primo tipo di problema nasce quando le osservazioni riguardano due gruppi ben distinti di unità e per i quali i valori medi delle variabili X e Y sono sufficientemente differenti, cioè quando siamo in presenza di due nuvole differenziate di punti.

Infatti, se per esempio il baricentro del primo gruppo è più in basso rispetto a quello del secondo gruppo, la retta ipotetica che attraversa la nuvola aggregata dei punti possiede una pendenza determinata dalla situazione relativa dei due baricentri, provocando un aumento «fittizio» nella correlazione. In tali casi, se si calcola la correlazione nei due gruppi separati si giunge ad un legame più basso di quello che si ottiene calcolando la correlazione sui gruppi aggregati.

Evidentemente, un effetto opposto si verifica quando la popolazione in esame si suddivide in sottogruppi i cui baricentri sono collocati «a forma di V», nel senso che in un primo gruppo vi è relazione decrescente tra X e Y mentre nel secondo vi è relazione crescente.

Ebbene, anche se nei due gruppi separatamente esaminati la correlazione è elevata in valore assoluto, verosimilmente, nel gruppo aggregato, si riscontrerà una modesta correlazione indotta dal fatto che la retta che attraversa tale nuvola di punti tenderà a disporsi parallelamente all'asse delle ascisse indicando, quindi, un'indipendenza in media di Y al variare di X.

La problematica della correlazione a blocchi si ritrova nell'analisi di dati antropometrici rispetto al sesso, ma anche per dati economici e sociali. Per esempio, in una realtà territorialmente diversificata come quella del nostro Paese, i centri medi di Redditi e Consumi sono molto distanti tra i gruppi residenti in territori differenti. Se si calcola la correlazione tra Redditi e Consumi di unità familiari, all'interno di gruppi distinti tra Nord e Sud del Paese si avrà una correlazione quasi certamente più bassa di quella che si ottiene aggregando la popolazione di riferimento.

- ***Correlazione spuria.*** Il secondo tipo di problema è così definito perché si manifesta quando esiste un legame tra X e Y solo perché entrambe le variabili sono funzioni di una terza variabile Z che condiziona entrambe.

La letteratura statistica registra numerosi esempi paradossali di tali situazioni:

i) la successione oraria dei prezzi di cavolini di Bruxelles venduti nel mercato di Londra e la corrispondente altezza del Tamigi rilevata ogni ora evidenziano una fortissima correlazione perché entrambi i fenomeni seguono un andamento paraboloide nell'arco di una giornata;

ii) il numero mensile di turisti canadesi che giungono a Roma e il numero di gelati venduti nella nostra capitale presentano una forte correlazione lineare perché la variabile che veramente determina il loro andamento è la temperatura le cui variazioni stagionali determinano sia il fenomeno degli arrivi di turisti canadesi che quello della vendita dei gelati; etc.

Tuttavia, in aggiunta a tali esempi, «agevoli da smascherare», il fenomeno della correlazione spuria è ben presente (per citare solo alcuni casi) nelle scienze economiche, sociali e sanitarie quando la ricerca di collegamenti tra le variabili del mondo reale è complessa a causa delle numerose interrelazioni, difficili da esaminare separatamente, come per esempio:

i) in Medicina, la complessità delle relazioni tra gli organi del corpo umano impone una elevata cautela nella diagnostica clinica, perché indicatori locali sono spesso modificabili in funzione di condizioni generali, vaghe e difficili da specificare;

ii) su dati socio-economici o ambientali, tali effetti sono particolarmente frequenti quando i dati si raccolgono come serie temporali, perché la comune evoluzione di due fenomeni dinamici è determinata dal comune contesto congiunturale o ambientale.

Un modo per affrontare correttamente tali problemi è quello del calcolo della *correlazione parziale* che richiede sviluppi analitici più complessi, e che non affronteremo in questo volume.

- ***Dati estremi.*** La presenza di dati estremi (anomali ma esistenti oppure semplicemente errati, e quindi da eliminare o correggere) modifica sicuramente le medie di X e/o Y, ma molto di più influisce sul prodotto degli scarti nel contribuire a sminuire – ed in casi particolari, sinanche a capovolgere – la forza e la direzione del legame lineare misurato dal coefficiente di correlazione ρ.

Anche se in tale ambito sono state proposte misure resistenti per il coefficiente di correlazione lineare, è decisamente più diffusa la prassi di eliminare e/o sostituire i dati estremi con valori più centrali (ottenuti mediante una ragionevole *imputazione*), prima di procedere al calcolo di ρ.

Ambiente R

Distribuzioni statistiche multiple

R7.1 Introduzione

Anche se il testo è sostanzialmente dedicato allo studio di una variabile statistica doppia (X, Y), grazie alle esemplificazioni dei data set, presenteremo anche elaborazioni con più di due variabili mediante l'uso di `R` .

Sequenzialmente, illustreremo le seguenti funzioni:

- Funzioni per ottenere da un *data frame* una tabella doppia di frequenze (assolute, relative, marginali).
- Funzioni per visualizzare relazioni tra variabili quantitative e variabili quantitative condizionate da fattori.
- Funzioni per il calcolo degli indici statistici di connessione e correlazione tra variabili.

R7.2 Distribuzioni doppie di frequenza

Nel caso di variabili con modalità elencabili (qualitative oppure quantitative discrete), la funzione per ottenere tabelle doppie di frequenze è `table()` come nell'esempio seguente. Qui esaminiamo la distribuzione della variabile doppia `(componenti,genere)` nel data set `studenti`, dove la prima componente è discreta e la seconda è qualitativa (dicotomica, indicata con 0 per gli studenti e 1 per le studentesse). Dividendo per la somma totale delle frequenze, si ottengono le frequenze relative (che arrotondiamo a 5 decimali), come nelle istruzioni seguenti:

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> genere=dati[,2]; componenti=dati[,10];
R==> tabella=table(componenti,genere);
R==> tabella;
           genere
componenti   0   1
         1  12   2
         2  26  29
         3 142 190
```

```
        4  422 635
        5  314 560
        6  117 187
        7   27  64
        8   16   8
        9    0   3
        10   2   2
        12   1   0
R==> round(tabella/length(genere), digits=5);
          genere
componenti       0       1
        1  0.00435 0.00072
        2  0.00942 0.01051
        3  0.05147 0.06887
        4  0.15295 0.23016
        5  0.11381 0.20297
        6  0.04241 0.06778
        7  0.00979 0.02320
        8  0.00580 0.00290
        9  0.00000 0.00109
        10 0.00072 0.00072
        12 0.00036 0.00000
```

Le distribuzioni marginali si ottengono con la funzione `margin.table(tabella,codice)`, dove `codice=1` per le righe e `codice=2` per le colonne, come esemplificato nelle istruzioni seguenti. Occorre poi dividere per il totale delle frequenze assolute (indicato con n) per ottenere le frequenze marginali relative.

```
### frequenze relative marginali di riga (variabile "componenti")
R==> round(margin.table(tabella,1)/n,digits=5);
componenti
      1       2       3       4       5       6
0.00507 0.01993 0.12033 0.38311 0.31678 0.11018
      7       8       9      10      12
0.03298 0.00870 0.00109 0.00145 0.00036

### frequenze relative marginali di colonna (variabile 2genere")
R==> round(margin.table(tabella,2)/n,digits=5);
genere
      0       1
0.39108 0.60892
```

Non va dimenticato che nel data set `studenti` nessuna unità statistica vive in una famiglia con 11 componenti, per cui tale modalità (che dovrebbe avere frequenze nulle in una tabella completa) non è considerata tra quelle ammissibili dalla funzione `table` di R perché non osservata nel data set. Invece, per una disamina completa delle modalità (da 1 al massimo) va considerata la funzione `tabulate` che, nello specifico, produce il seguente risultato:

```
R==> tabella=tabulate(componenti);
R==> round(tabella/n, digits=5);
 [1] 0.00507 0.01993 0.12033 0.38311 0.31678 0.11018 0.03298
 [8] 0.00870 0.001090.00145 0.00000 0.00036
```

che coincide con il precedente ma, come mostra il penultimo numero elencato, include anche la frequenza relativa nulla della modalità `componenti=11`.

Infine, ricordiamo che si possono anche ottenere direttamente le frequenze relative marginali (=*proportion*) utilizzando la funzione `prop.table` con modalità analoghe.

Quando una o entrambe le variabili componenti sono continue, è necessario suddividere in classi l'intervallo di variazione della variabile continua, applicando poi a tali variabili (di fatto discretizzate in classi) la funzione `table()` e `margin.table()` come in precedenza.

L'esempio illustra tale utilizzo per la variabile doppia continua (`altezza, peso`) nel data set `studenti`; si noti la funzione `range()` per determinare il campo di variazione dei dati prima di stabilire in quante e quali classi operare la suddivisione.

Quindi, si applica la funzione `cut(variabile, suddivisioni)` che costruisce dei codici per ciascuna modalità di variabile in funzione dell'intervallo in cui ricade, così come specificato da suddivisioni (breaks); in effetti, `cut()` associa un fattore a ciascuna modalità contando poi tali fattori mediante la funzione `table`.

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> peso=dati[,4]; altezza=dati[,5]; n=length(genere);
R==> range(peso);range(altezza);
[1]  40 120
[1] 146 196
R==> breakalt=c(140,150,160,170,180,190,200);
R==> breakpes=c(39.9,50,60,70,80,90,100,110,120);
R==> tabella=table(cut(altezza,breakalt),cut(peso,breakpes));
R==> tabella;
           (39.9,50] (50,60] (60,70] (70,80] (80,90]
(140,150]         8       4       0       0       0
(150,160]       195     280      31       6       1
(160,170]       112     687     327      59      11
(170,180]         2     108     298     299      72
(180,190]         0       2      21      95      80
(190,200]         0       0       0       4       5

           (90,100] (100,110] (110,120]
(140,150]        0         0         0
(150,160]        0         0         0
(160,170]        1         1         0
(170,180]        6         4         0
(180,190]       30         2         2
(190,200]        3         2         1
```

Ovviamente, le tabelle marginali e le frequenze assolute e relative si ottengono nel modo già evidenziato nel precedente esempio.

Si osservi la specificazione della prima classe con il valore 39.9 anziché 40 (che è l'effettivo peso minimo dichiarato): ciò è necessario perché – trattandosi di classi chiuse a destra ed aperte a sinistra – se si fosse specificato il minimo come 40, le unità statistiche con peso esattamente uguale a 40 sarebbero state escluse dal conteggio. Infine, le distribuzioni condizionate delle frequenze assolute si ottengono semplicemente considerando la riga (colonna) di interesse della distribuzione doppia; dividendo, poi, tale distribuzione univariata per il totale delle frequenze si ottengono le frequenze relative.

R7.3 Rappresentazioni grafiche di variabili doppie

Le rappresentazioni grafiche di una distribuzione doppia richiederebbero un istogramma bivariato (e questo non è standard in R) per cui è di maggiore interesse visualizzare il tipo di relazioni di queste variabili in funzione della loro natura.

Condizionamento di una variabile qualitativa

Per visualizzare una variabile quantitativa in funzione delle modalità di una variabile qualitativa (*factor*) si possono utilizzare i box-plots al variare delle modalità della seconda variabile. L'esempio illustra tale possibilità per la variabile `eta` (continua) in funzione della variabile `diploma` (qualitativa) nel data set `studenti`.

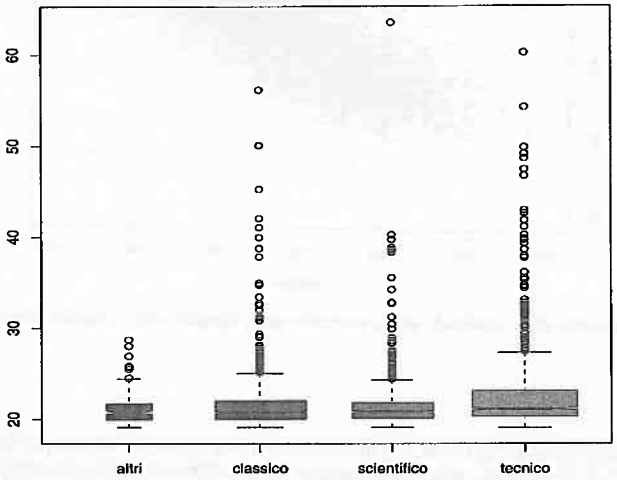

FIGURA R7.1. *Box-plot della variabile età, al variare del fattore diploma.*

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> attach(dati);
R==> boxplot(eta~diploma,varwidth=TRUE,notch=TRUE,
        col=gray(c(0.3,0.4,0.5,0.6)),ylab="Età",
        names=c("altri","classico","scientifico","tecnico"));
R==> detach(dati);
```

In effetti, la Figura R 7.1 evidenzia solo un leggero aumento dell'età mediana per i provenienti dal diploma tecnico rispetto agli altri studenti.

Scatter plot per una coppia di variabili continue

Quando (X,Y) è una variabile doppia continua, il grafico più utile è la "nuvola dei punti" (*scatter plot*) mediante il quale sugli assi cartesiani si rappresentano le coppie (x_i,y_i), $i=1,2,\ldots,n$. Quando esistono sottogruppi nella popolazione in esame, qualche opzione grafica permette di fatto anche di evidenziare l'effetto di una variabile qualitativa (fattore), come ora mostreremo.

In effetti, abbiamo già utilizzato più volte la funzione grafica primaria `plot()`, e ricordiamo che il grafico di y rispetto a x si può disegnare con una delle due istruzioni equivalenti:

```
R==> plot(x,y);    ### in ascissa x, in ordinata y
R==> plot(y~x);    ### grafico di y rispetto a x
```


Per esempio, utilizzando il data set `camellia`, la Figura R 7.2 presenta la variabile `pesosecco` in funzione di `pesovivo` accompagnata da una ipotetica relazione approssimata (che si approfondirà nel Capitolo 18) ed espressa mediante una retta passante per l'origine e sovrapposta allo *scatter plot*. Tale rappresentazione si ottiene con le istruzioni:

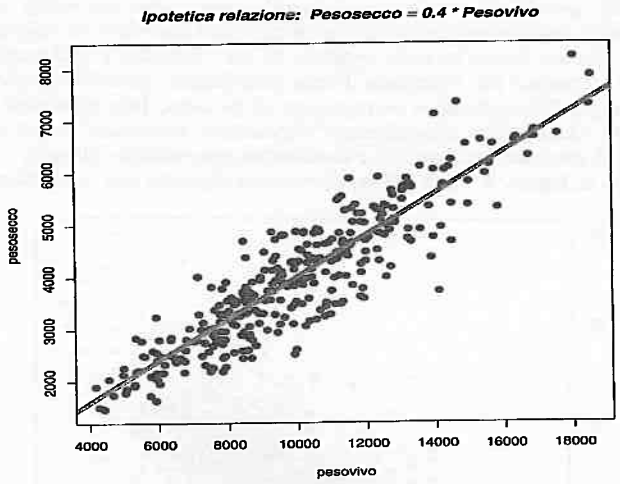

FIGURA R7.2. *Relazione tra pesosecco e pesovivo nel data set camellia.*

```
R==> dati=read.table("camellia.txt", header=TRUE);
R==> attach(dati);
R==> plot(pesosecco~pesovivo,pch=19,lwd=2);
R==> title(main="Ipotetica relazione:  Pesosecco =
        0.4 * Pesovivo",
        font.main=4);
R==> detach(dati);
R==> abline(0,0.4,col=gray(0.6),lwd=6);
```

L'ultima istruzione: `abline(0,0.4,col=gray(0.6),lwd=6);` utilizza la funzione grafica secondaria `abline()` (con opzioni di colore e spessore) la quale sovrappone una retta $y=a+bx$, con intercetta $a=0$ e coefficiente angolare $b=0.4$, espressi dai primi due valori della funzione.

Effetto *jittering* nelle rappresentazioni grafiche

Può accadere che la rappresentazione grafica in due dimensioni nasconda la natura delle relazioni tra variabili perché i punti di "sovrappongono", rendendo di fatto indistinguibile un punto isolato da un altro che, invece, si verifica con elevata frequenza. Questa situazione accade quando le modalità sono ordinali ovvero consistono in poche modalità discrete. A tale inconveniente si potrebbe rimediare con una rappresentazione in tre dimensioni ma questa non è sempre efficace (dipendendo dalla prospettiva, dal software disponibile, e così via).

Allora, anziché rappresentare i punti osservati (x_i, y_i) si rappresentano i punti (x_i^*, y_i^*) definiti da:

$$x_i^* = x_i + u_i\,; \qquad y_i^* = y_i + v_i$$

dove $(u_i\,, v_i)$ sono la realizzazione di due variabili casuali indipendenti di valore medio 0 e varianza molto piccola (da stabilire dopo qualche prova, caso per caso). Anche se il concetto di variabile casuale sarà introdotto a partire dal Capitolo 9, la sostanza di tale operazione è di alterare (come se fosse soggetto ad un "tremolio") ogni punto (x_i, y_i), modificandone sia l'ascissa che l'ordinata di una piccolissima quantità casuale in modo che sia di fatto impossibile che esso si sovrapponga ad un altro. Tale operazione si chiama *jittering* (da *jitter*, che significa letteralmente "nervosismo, agitazione" e, per estensione, "piccole, rapide ed irregolari variazioni") e si dimostra generalmente efficace.

Consideriamo la Figura R 7.1 del testo che è stata ottenuta con le istruzioni:

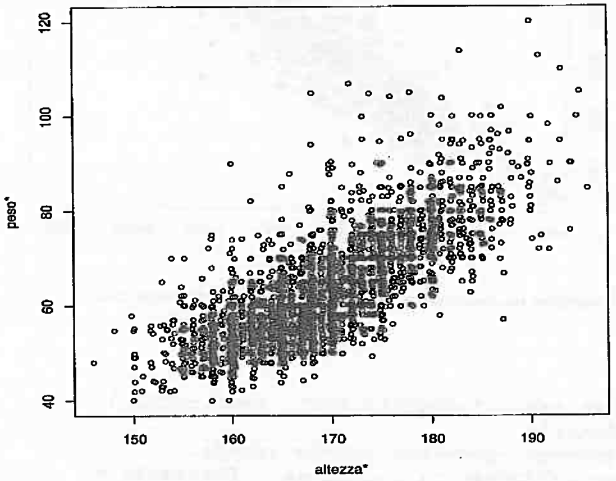

FIGURA R7.3. Jittering *per le variabili (altezza,peso) nel data set studenti.*

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> attach(dati);
R==> plot(peso~altezza,xlim=c(140,200),ylim=c(20,120),pch=20);
R==> detach(dati);
```

A partire da essa, le istruzioni seguenti costruiscono un grafico che applica la tecnica *jittering* sia alla variabile peso che altezza nel data set studenti: tali variabili sono dichiarate dagli intervistati (e non misurate) per cui risentono degli arrotondamenti che avvengono nella dichiarazione non solo all'intero più vicino (questo sarebbe casuale) ma ai multipli di 5.

```
dati=read.table("studenti.txt", header=TRUE);
peso=dati[,4]; altezza=dati[,5]; n=length(altezza);
sigma=0.15;
pesostar=peso+sigma*rnorm(n); altezzastar=altezza+sigma*rnorm(n);
plot(altezzastar,pesostar,pch=21,cex=0.75,
            xlab="altezza*",ylab="peso*")
```


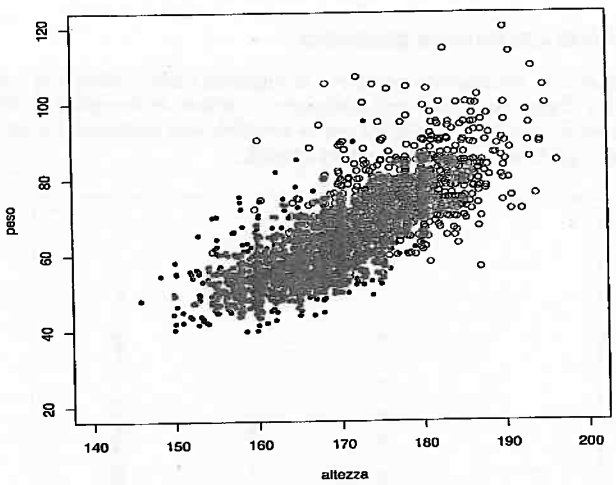

FIGURA R7.4. Scatter plot *(*jittered*) della variabile peso rispetto alla variabile altezza, distinta per genere.*

Ebbene, se si confronta la Figura R 7.1 del testo con la Figura R 7.3 qui prodotta emerge con chiarezza come lo *jittering* abbia evidenziato degli addensamenti sui multipli di 5 nelle dichiarazioni delle misure di altezza e peso, molto pronunciate fra 160 e 180 cm e fra 60 e 80 kg, rispettivamente. Per distinguere anche il genere si possono usare opzioni colorate ovvero simboli differenti per studenti e studentesse. Inoltre, poiché (per l'arrotondamento agli interi per cm e kg dichiarati) molti dati si replicano è opportuno alterare leggermente ed in modo casuale ascissa ed ordinata di ogni punto mediante la tecnica *jittering*. Alla fine, la Figura R 7.4 è stata ottenuta con le seguenti istruzioni:

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> genere=dati[,2]; n=length(genere);
R==> peso=dati[,4]; altezza=dati[,5];

R==> colgen=ifelse(genere==1,"red","blue"); ### differenti colori
R==> pchgen=ifelse(genere==1,20,21);        ### differenti simboli
R==> altezza=altezza+0.25*rnorm(n);         ### altezza jittered
R==> peso=peso+0.25*rnorm(n);               ### peso jittered

R==> plot(peso~altezza,xlim=c(140,200),ylim=c(20,120),
          col=colgen,pch=pchgen);
```

Il lettore apprezzerà maggiormente sul proprio schermo grafico a colori l'effetto prodotto.

Scatter plots per variabili multivariate

Quando si hanno più variabili quantitative, una tecnica efficace è quella di disegnare gli *scatter plot* a coppie, e questo si ottiene con la funzione pairs() applicata ad una lista di vettori. Per esempio, utilizzando il data set studenti si può rappresentare la variabile tripla (*peso, altezza, eta*) ed ottenere la Figura R 7.5 con le istruzioni seguenti:

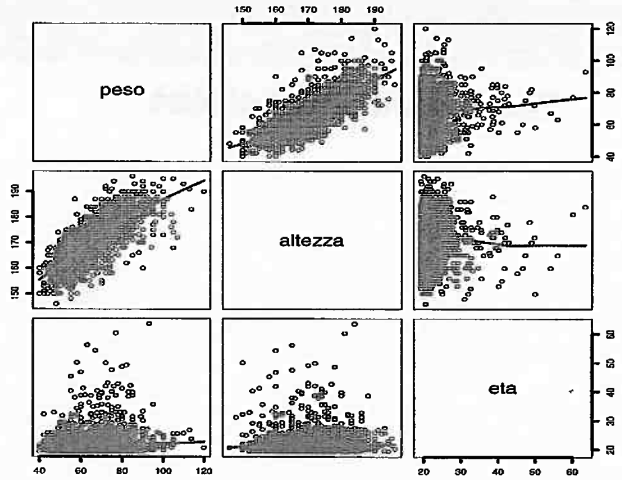

FIGURA R7.5. *Coppie di* scatter plot *per la variabile tripla (peso,altezza,eta).*

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> peso=dati[,4]; altezza=dati[,5]; eta=dati[,3];

R==> pairs(cbind(peso,altezza,eta),panel=panel.smooth,lw=3);
```

L'opzione `panel=panel.smooth` ha sovrapposto alle nuvole dei punti di ciascun riquadro una curva regolare che interpola al meglio la linea di tendenza tra ordinate ed ascisse e, quindi, mostra se ed in qual modo le variabili sono legate: la determinazione di tale linea esula, però, dal livello di questo testo.

Le potenzialità di R sono molto estese sia con altre specifiche opzioni della funzione `pairs` che utilizzando altri pacchetti grafici più specializzati (come `grid` e `lattice`, per esempio).

R7.4 Misure di connessione

Anche se per calcolare l'indice X^2 di Pearson occorre confrontare frequenze osservate e frequenze teoriche rispetto alla ipotesi estrema di indipendenza di due qualsiasi variabili, X e Y, la formula più comoda per le elaborazioni è quella che utilizza le sole frequenze osservate (assolute e marginali). Pertanto, utilizzando la simbologia standard, come riportato nel testo:

$$X^2 = n\left(\sum_{i=1}^{k}\sum_{j}^{h}\frac{n_{ij}^2}{n_{i\cdot}\,n_{\cdot j}} - 1\right).$$

A tal fine, abbiamo costruito la seguente funzione `X2()` che si applica ad un oggetto `tabella` (frequenze assolute) ottenuta con la funzione `table()`. In essa abbiamo utilizzato la funzione `outer()` rispetto alla operazione prodotto (*) che effettua il prodotto

cartesiano – cioè tutti i possibili prodotti a due a due – tra gli argomenti della funzione (che, qui, sono le frequenze marginali).

```
R==> X2<-function(tabella){
          ni=margin.table(tabella,1);
          nj=margin.table(tabella,2);
          denom=outer(ni,nj,"*");
          sum(tabella)*(sum((tabella^2)/denom)-1);
          }
```

Infine, l'indice normalizzato di Cramér si ottiene mediante la seguente funzione `Cramer`:

```
R==> Cramer<-function(tabella){
              sqrt(X2(tabella)/(sum(tabella)*min(dim(tabella)-1))
          }
```

Essa è un esempio di funzione definita tramite un'altra funzione (cioè, `X2()` in questo caso).

Esemplificando, per misurare la connessione fra le variabili `diploma` e `genere` nel data set `studenti`, si utilizzano le seguenti istruzioni:

```
R==> dati=read.table("studenti.txt", header=TRUE);
R==> genere=dati[,2]; diploma=dati[,13];
R==> tabella=table(genere,diploma);
R==> tabella;
       diploma
genere alt cla sci tec
     0  29 213 308 529
     1 182 608 329 561
R==> X2(tabella);Cramer(tabella);
[1] 180.25
[1] 0.25560
```

Esiste quindi una certa connessione (pari al 26% di quella massima) tra le variabili qualitative `diploma` e `genere`. Si osservi che la letteratura riporta che l'indice di Cramér raggiunge il suo massimo solo in casi estremi, per cui il valore trovato indica una connessione tutt'altro che trascurabile anche se numericamente modesta.

R7.5 Correlazione tra variabili statistiche

Il coefficiente di correlazione tra due variabili quantitative X e Y si calcola in R mediante la funzione `cor()` che può essere applicata a tutte le colonne di una matrice di variabili quantitative oppure ad una selezione di variabili. Per esempio, per tutte le variabili presenti nel data set `camellia`, si ottiene una matrice di correlazione che all'elemento di posto (i,j) presenta il coefficiente di correlazione lineare tra le variabili X_i e X_j, cioè $Corr(X_i, X_j)$.

```
R==> dati=read.table("camellia.txt", header=TRUE);
R==> dati=dati[,2:6];
R==> round(cor(dati),digits=3);
          lunghezza larghezza pesovivo pesosecco codice
lunghezza     1.000     0.524    0.676     0.639   0.293
larghezza     0.524     1.000    0.764     0.597  -0.356
pesovivo      0.676     0.764    1.000     0.907  -0.111
pesosecco     0.639     0.597    0.907     1.000   0.001
codice        0.293    -0.356   -0.111     0.001   1.000
```

Dal data set è stata esclusa la prima colonna poiché essa contiene l'identificativo della foglia (e non ha, quindi, alcun significato statistico). Si evidenzia come il massimo legame lineare (positivo) sia tra le variabili `pesovivo` e `pesosecco`. Per interpretare anche la correlazione misurata con la variabile dicotomica `codice` (=1 se la specie è Lanceolata), si osservi che la correlazione è positiva con `lunghezza` (il che è logico, essendo la specie Lanceolata dotata di foglie più appuntite) ed è negativa con `pesovivo` (perché le foglie più appuntite sono tendenzialmente più leggere rispetto a quelle della specie Cordiforme).

Parte seconda

Teoria della probabilità

Capitolo ottavo

Calcolo delle probabilità

8.1 Introduzione

La variabilità presente nella vita quotidiana genera incertezza nelle decisioni da assumere e induce l'Uomo a ricercare comportamenti razionali, anche quando non è sicuro dei risultati delle azioni che intraprende, soprattutto rispetto al futuro: da questa immanente insicurezza derivano il bisogno e l'utilità di studiare la probabilità.

In effetti, l'osservazione dei fenomeni naturali (fisici, chimici, biologici, meteorologici, etc.) e sociali (economici, finanziari, psicologici, politici, etc.) mostra che la previsione accurata di un risultato è resa difficile, se non impossibile, da meccanismi molto complessi; anzi, man mano che si approfondisce lo studio di un fenomeno, la indeterminatezza nella previsione del risultato aumenta.

- Un sasso lanciato in aria cade sempre; ma, se si studia la sua velocità di caduta o si cerca di stabilire il punto esatto in cui ricadrà al suolo, la risposta diventa tutt'altro che univoca. Anche ripetendo l'esperimento con la massima accuratezza, difficilmente si avrà lo stesso risultato.
- Nel tiro con l'arco si cerca di colpire il centro di un bersaglio ma anche chi si allena per molto tempo difficilmente può escludere un margine di errore nei vari tiri. Anche il tiratore più esperto difficilmente potrà conoscere in anticipo l'esatto punto in cui la freccia colpirà il bersaglio.
- Le previsioni meteorologiche «locali» sono tra gli eventi più difficili da prevedere e solo l'introduzione dei satelliti ha ridotto il numero di insuccessi rispetto alle previsioni del passato. Eppure, l'insicurezza nella conoscenza della temperatura di domani non impedisce alle persone di vivere con relativa tranquillità, di vestirsi quasi sempre adeguatamente e persino di pianificare iniziative all'aperto senza eccessivi traumi. Ciò avviene perché il grado di precisione necessario per la vita quotidiana è piuttosto grossolano in rapporto a quello imposto per il lancio di una capsula spaziale (ove la variazione di qualche grado centigrado nella temperatura può creare notevoli problemi).
- Ogni politico percepisce, con una certa accuratezza, la percentuale dei consensi che il suo partito riceverà alle prossime elezioni, ma nessuno seriamente potrà asserire di conoscerla in anticipo con esattezza.

In tali ambiti, è tuttora aperta la discussione circa il ruolo dell'indetermini-

smo nella conoscenza umana: è colpa della nostra ignoranza nel valutare tutte le cause che, in numero enorme, determinano gli effetti visibili, oppure esiste un limite oggettivo per cui è impossibile aumentare il livello di accuratezza nella conoscenza senza incontrare l'indeterminismo? La ricerca scientifica moderna è orientata verso la seconda ipotesi, anche se molti scienziati – che pure hanno apportato alla interpretazione probabilistica contributi essenziali (da Laplace a Einstein) – hanno avuto difficoltà ad accettare una concezione essenzialmente indeterministica.

Motivi storici ed esigenze didattiche hanno spesso associato il Calcolo delle probabilità ai giochi a sorte: in tali casi, le situazioni sono ben schematizzabili, gli esiti sono elencabili con precisione e le aspettative sono valutabili secondo il comune «buon senso». Tuttavia, l'utilità e l'efficacia di una Teoria della probabilità si manifesta proprio in quelle situazioni nelle quali ci si allontana il più possibile dai giochi a sorte, come avviene regolarmente nella ricerca scientifica e nelle applicazioni delle scienze naturali e socio-economiche.

Questo Capitolo offre una presentazione ormai ben consolidata del Calcolo delle probabilità che si è dimostrata efficace e coerente, soprattutto per le applicazioni statistiche.

8.2 Concetto e misura della probabilità

È importante esaminare la distinzione tra *concetto* e *misura* della probabilità, poiché essa evidenzia il lungo percorso intellettuale che ha condotto all'attuale formalizzazione.

- In primo luogo, ***la probabilità è un concetto primitivo***, cioè originario per l'essere umano perché innato e sempre presente nelle sue regole di comportamento. In quanto tale, esso obbedisce a criteri logici coerenti traducibili in un sistema di assiomi dai quali si possono, con deduzioni rigorose, dimostrare i teoremi.

- In secondo luogo, ***la probabilità è una misura*** perché associa al concetto primitivo una valutazione numerica. Tuttavia, andrebbe sempre distinto il momento della percezione concettuale da quello della valutazione numerica.

○ Una persona adulta, priva di qualsiasi nozione di Calcolo delle probabilità, intuisce perfettamente che la vincita al Lotto di una quaterna è molto meno probabile di quella di un ambo, e similmente che tre giorni consecutivi di pioggia a Napoli dal 13 al 15 agosto sono molto meno probabili che dal 23 al 25 dicembre. Eppure, poche persone saprebbero determinare con immediatezza la *misura* di queste probabilità, anche se il *concetto* di probabilità viene percepito e graduato anche da chi non possiede alcuna conoscenza statistica.

Vediamo, allora, cosa caratterizza un esperimento probabilistico e la percezione che noi abbiamo circa i possibili esiti di una prova.

8.3 Caratteristiche degli esperimenti probabilistici

I fenomeni di cui si occupa il Calcolo delle probabilità riguardano le scienze fisiche, naturali, umane e forniscono esiti così differenti tra loro che una elencazione esaustiva appare impossibile. Ciascun lettore può ampliare la gamma delle possibili esemplificazioni.

○ Per motivi molto diversi, si può essere interessati al presentarsi a Roma di un giorno sereno dopo due nuvolosi; all'uscita della faccia con 3 punti nel lancio di un dado; alla sopravvivenza di una persona oltre i 100 anni; al presentarsi dei numeri 1, 2, 3, 4, 5 sulla ruota di Firenze nella prossima estrazione del Lotto; al contenimento dell'inflazione per il 2005 tra l'1.5% e l'1.8%; al verificarsi di una nuova varietà di gerani tra 100 semi accuratamente impollinati; al giocare a carte sino a che uno dei due giocatori non perde tutto il capitale iniziale; al superamento di 200 ore per il funzionamento di una lampadina elettrica; alla vincita di un premio superiore a 1 milione di Euro nella lotteria Italia; al superare l'esame di Statistica avendo studiato solo il 75% del programma; al classificarsi primo ad una competizione sportiva; all'indovinare in quale giorno della settimana è nato Manzoni; al presentarsi di 7 pareggi nella prossima schedina del Totocalcio; all'acquisto di una confezione di dentifricio reclamizzato nell'ultima settimana da parte di un giovane che entra in un supermercato; alla velocità di un elettrone di un atomo di ossigeno; all'essere realmente portatori di una infezione se un test clinico l'ha segnalata; al tempo impiegato per raggiungere l'ufficio in auto questa mattina; all'indovinare 8 cifre sulle 10 che una persona ha segretamente scritto su di un foglio, e così via.

Una accurata riflessione sugli esperimenti appena elencati mostra che tra loro sussistono alcuni elementi in comune e che questi sono presenti con differente intensità. Essi sono: *la incertezza del risultato, la ripetibilità dell'esperimento, la equiprobabilità dei risultati.*

- In primo luogo, la ***incertezza del risultato*** deriva dal fatto che nelle situazioni esaminate sono possibili più esiti. Tali possibili risultati sono elencabili (come, per esempio: le sei facce del dado, i possibili voti in trentesimi di un esame universitario, etc.) oppure possono essere concettualmente assimilati a tutti i numeri inclusi in un intervallo reale (per esempio: il tasso di inflazione dell'anno 2005, la durata di funzionamento di una lampadina elettrica, il tempo occorrente per percorrere con un'auto la distanza di 10 km, etc.).

Poiché l'incertezza è generata da meccanismi molto diversi, essa presenta anche aspetti sostanzialmente differenti. Infatti, l'incertezza di un esperimento può essere:

i) *momentanea*: domani sera saprò con certezza quale tempo avrà fatto domani; se consulto una enciclopedia e svolgo qualche calcolo conoscerò perfettamente il giorno della settimana in cui è nato Manzoni;

ii) *tecnologica*: fino a qualche decennio fa, conoscere anzitempo il sesso dei nascituri era quasi una divinazione, così come discutere della composizione del suolo lunare;

iii) *fisica*: in linea di principio, come sosteneva Laplace, potrei elencare tutte le forze che imprimo alla moneta nel lanciarla in aria e formalizzare in equazioni matematiche le sue interazioni con l'aria, la gravità, l'impatto con il terreno, e così via: allora, se fossi in grado di risolvere le tantissime equazioni differenziali così impostate sarei anche in grado di conoscere, supposte note le condizioni iniziali, quale sarà

la posizione finale della moneta. L'impossibilità pratica di impostare e risolvere in tempi ragionevoli questo problema genererebbe l'incertezza nel risultato;

iv) *intrinseca*: il principio di indeterminismo enunciato nel 1927 da Heisenberg per la Fisica assicura che è impossibile determinare contemporaneamente posizione e velocità di una particella;

v) *psicologica*: grazie alla libertà individuale e all'imponderabile che presiede a tutte le decisioni umane, non saprò mai prevederle in anticipo con assoluta certezza.

• In secondo luogo, la ***ripetibilità dell'esperimento*** rende lecito chiedersi se quell'esito si verificherà nuovamente in circostanze diverse o anche simili. Quest'affermazione è molto delicata perché, in senso stretto, *un esperimento reale è sempre fisicamente irripetibile*, perché tutto si modifica nell'Universo, rispetto allo spazio e/o al tempo. Tuttavia, tale rigore concettuale è raramente utile perché l'esperienza umana si accresce grazie alla similarità e non alla perfetta identità tra circostanze simili. Quindi, molti esperimenti possono essere considerati «come se» fossero stati ripetuti nelle medesime condizioni in rapporto a quegli elementi essenziali che interessano lo studioso, lo scienziato o il giocatore.

i) Il lancio di una moneta viene considerato un esperimento ripetibile all'infinito se si ha l'accortezza di lanciarla sufficientemente in alto e di fare in modo che essa, cadendo, possa liberamente rotolare sino a posarsi su una delle due facce. In questo senso, poiché le monete dell'epoca moderna sono sempre circolari e simmetriche, con fregi e scritte che non possono alterare sensibilmente la simmetria delle facce, la maggior parte delle persone riterrà indifferente giocare a «Testa e Croce» utilizzando monete da 50 centesimi, 1 o 2 Euro.

ii) L'estrazione ripetuta di una pallina da un'urna che contiene palline sferiche è un esperimento ripetibile se tali palline sono costruite con lo stesso materiale e possiedono stesso peso, rugosità e temperatura al punto da renderle indistinguibili rispetto all'estrazione di una di esse, se accuratamente mescolate.

iii) Lo studio costante e ripetuto nel tempo della durata di funzionamento di una lampadina, prescelta fra alcune migliaia, non dipenderà dal numero di serie della lampadina prescelta, se si ipotizza che nel processo di produzione delle medesime non sono variate apprezzabilmente le condizioni che modificano la resistenza del filamento. Quindi, la registrazione ripetuta della durata di funzionamento di ciascuna di esse si può assimilare ad un esperimento ripetuto nelle medesime condizioni.

Non tutti gli esperimenti sono ripetibili, per definizione o per ragioni strutturali. Il primo trapianto cardiaco e il primo allunaggio costituiscono esempi dei quali non si avevano esperienze precedenti; accurate prove su animali da laboratori, nel primo caso, e simulazioni sotto condizioni analoghe sulla Terra, nel secondo caso, sono tentativi di ridurre i rischi di esperimenti unici, assimilandoli ad altri che si suppongono simili ma ripetibili.

• Infine, la ***equiprobabilità dei risultati*** esprime la indifferenza di aspettativa rispetto al verificarsi di ciascuno degli esiti possibili, nel senso che non esistono motivi per attribuire maggiore fiducia al presentarsi di un risultato anziché di un altro. Spesso, la equiprobabilità è considerata un elemento di equità nei giochi di azzardo perché rende uguali tutti gli scommettitori. In ultima analisi, tale concetto deriva sia da un principio di indifferenza che di «ragione insufficiente»: non si vedono, cioè, ragioni sufficienti per attribuire ad uno dei possibili risultati maggiori opportunità rispetto agli altri. Sul piano pratico, ciò deriva da argomentazioni di simmetria fisica, geometrica o psicologica, come gli esempi confermano.

o La «equivalenza» tra banco e giocatore non sussiste nei giochi gestiti da privati (come nei vari casinò) o dallo Stato (come nel gioco del Lotto o del Totocalcio, in Italia), perché l'organizzatore – cioè il banco – deve comunque garantirsi un utile ed un margine economico per le spese ed i rischi connessi. Nel caso dello Stato, poi, il gioco regolamentato costituisce, di fatto, una imposta indiretta di natura volontaria su quei soggetti della collettività che lo praticano. Tali giochi sono tutt'altro che «equi» ma continuano a sussistere perché, nell'immaginario collettivo, assume rilevanza la sproporzione tra la modestia della posta giocata e il valore enorme della vincita, anche se per il singolo giocatore questa eventualità è rarissima.

i) Nel lancio di un dado simmetrico, composto da 6 facce quadrate, è ragionevole attribuire a ciascuna faccia una probabilità costante perché la omogeneità del materiale con cui è costruito il dado e le modalità del lancio non privilegiano una faccia più delle altre.

ii) Nell'estrazione di una pallina da un'urna che ne contiene 90, tutte di forma sferica e perfettamente omogenee rispetto alle caratteristiche fisiche e al cui interno ci sono biglietti numerati da 1 a 90, nulla può indurre a credere che colui che opera la scelta della pallina sia in grado di eseguire tecniche particolari per determinare un prefissato numero, essendo tutte di uguale dimensione, peso, rugosità, temperatura e, per giunta, accuratamente rimescolate prima di ogni nuova estrazione.

D'altra parte, la equiprobabilità dei risultati non può caratterizzare tutti gli esperimenti probabilistici sopra elencati perché, al di fuori dei giochi a sorte e dei meccanismi accuratamente preparati per studiare modelli probabilistici uniformi (l'esempio 10.1 dell'ago ruotante, di cui al paragrafo 10.2, è uno di essi), è piuttosto raro trovare situazioni reali che generano risultati differenti con la stessa probabilità. Infatti, ciascuno si sforza di accrescere la probabilità degli eventi «piacevoli» e di ridurre quella degli eventi «spiacevoli» per cui gli esperimenti reali presentano, spesso, situazioni tutt'altro che equiprobabili.

a) Nessuno studente che deve sostenere l'esame di Statistica ritiene equiprobabili i voti: 18, 19, ..., 30; infatti, conoscendo bene la sua preparazione, spesso più del docente, attribuisce – anche inconsciamente – probabilità diverse ai possibili voti.

b) Similmente, nessun agricoltore riterrà che il suo frutteto gli fornirà da 10 a 15 quintali di raccolto, da 15 a 20 quintali, da 20 a 25 quintali, ..., e così via, con la stessa probabilità.

c) In modo ancora più netto, la probabilità di vivere per i prossimi 5 anni per una persona sana è molto diversa se egli ha $20, 25, 30, \ldots, 90, 95, \ldots$ anni.

La discussione e gli esempi precedenti hanno mostrato in che senso *l'incertezza del risultato, la ripetibilità dell'esperimento e la equiprobabilità dei risultati* possono caratterizzare gli esperimenti probabilistici. Tuttavia, questi elementi non sono equivalenti. Infatti,

1. *tutti gli esperimenti sono incerti*;

2. *molti esperimenti sono riproducibili*;

3. *solo alcuni esperimenti danno luogo ad esiti equiprobabili.*

8.4 Concezioni alternative della probabilità

Le differenti interpretazioni che, storicamente, sono state fornite per la comprensione e l'utilizzo della probabilità hanno riguardato anzitutto gli esperimenti incerti, riproducibili ed equiprobabili; quindi, si sono rivolti a quelli incerti e riproducibili; infine, ai soli esperimenti soggetti ad incertezza, essendo quest'ultima l'unica caratteristica che richiede la introduzione di una probabilità.

o Le corrispondenti impostazioni sono state chiamate classica, frequentista, soggettivista e sono state proposte per giustificare come si forma nell'intelletto umano la misura della probabilità in contesti sempre più generali. Se, invece, cercano di definire la probabilità, tali proposte diventano tautologiche perché la probabilità è un concetto primitivo: quindi, non è definibile.

Per motivi didattici, la discussione che segue è molto schematica e di ciascuna impostazione esamineremo prima le ragioni che la rendono utile e poi le obiezioni critiche che le sono state rivolte. Ovviamente, ciò non rende giustizia ad un dibattito molto più ampio, che è risultato importante per lo sviluppo della conoscenza scientifica e che ha coinvolto filosofi della scienza, logici, epistemologi, matematici, fisici, probabilisti e statistici.

La nascita del Calcolo delle probabilità avviene con lo studio dei giochi di azzardo: prima utilizzando monete e dadi, poi con carte da gioco, roulette, Lotto e lotterie, scommesse sportive, etc. Per questo, è stato abbastanza naturale introdurre la ***definizione classica di probabilità*** – utilizzata di fatto da Galileo, Pascal, Fermat e posta per iscritto da Cardano ma ben formalizzata da Laplace solo nel 1812 – che è la seguente:

Dato un esperimento ben specificato ed un evento E tra quelli possibili per quell'esperimento, se m è il numero dei possibili risultati che danno luogo all'evento E mentre n è il numero di tutti i possibili risultati dell'esperimento, allora la probabilità dell'evento E è il rapporto m/n, purché tutti gli n risultati siano ugualmente possibili. Scriveremo, in breve: $P_r(E) = m/n$.

Gli esempi più semplici per tale definizione sono quelli relativi al lancio di una moneta o di un dado.

o I possibili risultati dell'esperimento «lancio di una moneta ben equilibrata» sono due: $T = Testa$ e $C = Croce$, e 1 solo tra i 2 possibili fornisce il risultato «si è verificata Testa» per cui, in coerenza con la definizione classica – se i due possibili risultati sono *ugualmente possibili* – la probabilità che esca Testa è: $P_r(T) = 1/2$.

o Similmente, nella prova «lancio di un dado non truccato e perfettamente simmetrico», i possibili risultati sono sei: $\{E_1, E_2, E_3, E_4, E_5, E_6\}$ avendo indicato con E_i il risultato per cui il dado si è fermato in modo che la faccia rivolta verso l'alto mostri i punti, per $i = 1, 2, \ldots, 6$. Allora, la probabilità della faccia con 3 punti sarà: $P_r(E_3) = 1/6$, perché vi è un solo esito E_3 per l'evento richiesto rispetto ai 6 *ugualmente possibili.*

La definizione classica della probabilità presenta l'indubbio vantaggio di calcolare la probabilità di un evento semplicemente come il rapporto tra la cardinalità (in pratica, la numerosità) di due insiemi: quello che interessa e quello che include la totalità dei possibili risultati dell'esperimento.

o Tale concezione si applica utilmente nella teoria della simulazione, nel campionamento da popolazioni finite, in alcuni modelli della fisica delle particelle, e così via. Ma essa presenta anche numerosi limiti.

Anzitutto, il concetto e la misura della probabilità non si possono circoscrivere ai giochi con esiti equiprobabili: anche ai dadi truccati i giocatori attribuiscono delle probabilità e nelle corse dei cavalli giocano poste differenziate in funzione della bravura (attesa o presunta) dei concorrenti.

In secondo luogo, esistono prove i cui eventi sono in numero infinito (numerabile oppure no) per i quali ha senso ragionare in termini di probabilità, ma per le quali la definizione classica diviene improponibile non essendo lecito il rapporto tra cardinalità infinite.

Infine, la definizione classica è tautologica perché per definire la probabilità occorre presupporre che gli esiti siano «tutti egualmente possibili» (=tutti ugualmente probabili), cioè *per definire la probabilità si suppone di conoscere il concetto di equiprobabilità* degli eventi.

Per rispondere a tali difficoltà, grazie soprattutto al lavoro di von Mises, la probabilità è stata formalizzata sulla scorta di una legge per eventi ripetibili, rispetto ai quali ciascun soggetto possiede una lunga e consolidata evidenza empirica: tale concezione della probabilità è molto antica e trova il sostegno in importanti risultati teorici (come la legge dei grandi numeri, di cui discuteremo nel paragrafo 11.2).

In pratica, la ***definizione frequentista della probabilità*** è la seguente:

Dato un esperimento ben specificato e perfettamente ripetibile, sia E un evento tra quelli possibili, ed indichiamo con $freq(E)$ il numero di volte che E si è verificato in una serie di n esperimenti ripetuti tutti nelle medesime condizioni, per cui $freq(E)/n$ è la frequenza relativa di E. Allora, la probabilità di E è il limite cui tende la frequenza relativa dell'evento E al crescere del numero degli esperimenti. Scriveremo, in breve: $P_r(E) = \lim_{n \to +\infty} \dfrac{freq(E)}{n}$.

o Evidenziamo prima gli aspetti positivi di tale presunta «definizione». Essa risponde ad un'esigenza operativa che include anche la situazione di eventi non equiprobabili, purché perfettamente riproducibili.

La probabilità viene misurata ripetendo indefinitamente l'esperimento e contando il numero delle volte che l'evento si è verificato in rapporto al totale (la frequenza relativa, appunto) nella convinzione (che deriva dall'esperienza) che tale numero si stabilizzerà al crescere del numero delle prove.

Sulla base di tale convincimento, è nato il concetto di mutualità tra i soggetti di una stessa collettività e, nei tempi moderni, quello di assicurazione. Infatti, constatando che nel lungo periodo il rapporto tra il numero di eventi «estremi» (naufragio delle navi, malattia, morte, etc.) e il numero dei soggetti a rischio (tutte le navi presenti in mare, la popolazione di un territorio, etc.) si stabilizzava, si è pensato di contribuire tutti in modo modesto per far fronte alle esigenze costose dei singoli individui colpiti da disgrazia.

Similmente, l'Astronomia ha sviluppato modelli previsivi sulla base delle osservazioni ripetute. La Demografia ha verificato la sostanziale stabilità del rapporto dei sessi alla nascita molto prima di pervenire a quelle conoscenze della Genetica che giustificano scientificamente tale regolarità (anche sulla base di modelli probabilistici). La ricerca

farmacologica è, in buona parte, fondata sulla sperimentazione ripetuta e selettiva per ottenere sostanze medicamentose sempre più efficaci sul piano terapeutico.

o Una critica importante deriva dal fatto che non è possibile definire analiticamente il «limite» implicito nella definizione frequentista. Tale «limite» non è paragonabile all'analogo concetto matematico perché si tratta solo di una constatazione empirica che nasce a posteriori dopo un gran numero di esperimenti.

o Una seconda critica riguarda il fatto che non tutti gli esperimenti sono ripetibili. Per esempio, è legittimo chiedersi quale sia la probabilità che:

– in altri sistemi solari si sia sviluppata una biologia paragonabile a quella terrestre;

– la dimostrazione dell'ultimo teorema di Fermat ottenuta da Wise nel 1995 sia corretta;

– entro il 2020 il tasso di riproduzione del continente africano sarà metà di quello attuale;

– i Cinesi raggiunsero le Americhe prima del 2000 a.C.;

– nel prossimo campionato di calcio le prime tre squadre si distanzieranno di 1 solo punto.

Eppure, in nessuno di questi casi, è possibile applicare la definizione frequentista di probabilità, né ovviamente quella classica.

Esistono, quindi, delle situazioni (con riguardo ad eventi passati, presenti e futuri) che richiedono valutazioni probabilistiche per le quali non è corretto riferirsi a schemi equiprobabili (ed applicare la definizione classica) né appellarsi alla ripetibilità dell'esperimento (ed applicare la definizione frequentista).

Questa insoddisfazione ha prodotto, nei decenni più recenti, soprattutto per merito di Ramsey, Savage e de Finetti, una impostazione alternativa della probabilità definita soggettivista (talvolta, si dice anche «soggettiva») perché la valutazione numerica della probabilità è connessa alle valutazioni personali di un soggetto circa la verificabilità di un esito incerto.

In pratica, la ***definizione soggettivista della probabilità*** è la seguente:

Dato un esperimento ben specificato, sia E un evento tra quelli possibili. Allora, la probabilità dell'evento E è la somma che un individuo coerente è disposto a scommettere in un gioco equo nel quale al verificarsi di E egli riceve dal banco un importo unitario.

Prima di discutere aspetti positivi e aspetti criticabili della definizione, è necessario specificare con alcuni esempi il senso di questa impostazione.

Perché la definizione non crei contraddizioni, il giocatore e il banco devono assumere un atteggiamento *coerente* nel senso che non possono accettare scommesse «contraddittorie» per le quali certamente uno di loro perderà o vincerà. Quindi, il gioco deve essere regolato in modo che per ciascuno sia indifferente assumere le parti del «banco» o del «giocatore».

o Nel gioco del lancio di una moneta ben equilibrata, ciascuno ritiene coerente ed equo scommettere 50 centesimi di Euro sulla faccia Testa per ricevere 1 Euro nel caso l'evento «Testa» si verifichi; tale ragionamento implica che: $Pr\,(Testa) = 1/2$.

Ciò che rende attraente la definizione soggettivista (rispetto alle due precedenti) è il fatto di non richiedere né la equiprobabilità degli eventi né la ripetibilità dell'esperimento. A volte, la definizione viene anche resa con la seguente frase: *La probabilità è la misura del grado di fiducia che una persona coerente attribuisce al verificarsi di un evento.*

o La critica più ricorrente alla concezione soggettivista è che essa produrrebbe tanti numeri per quanti sono gli individui di fronte allo stesso esperimento; tuttavia, dal ragionamento precedente, segue che se più individui «coerenti» sono disposti a scommettere cifre diverse essi stanno interpretando in modo differente quell'esperimento che, apparentemente, sembra identico.

o Più fondata, invece, ci sembra l'obiezione circa l'*origine* della valutazione numerica della probabilità per un soggettivista: proprio perché coerente, egli deriva tale misura dal ragionamento e/o dall'esperienza. Ma questo ricondurrebbe l'approccio soggettivista ad una generalizzazione dei precedenti approcci, rispettivamente, classico e frequentista.

o Infine, va sottolineato come nella concezione soggettivista di probabilità sia implicita un'idea di utilità che condiziona il concetto di probabilità. Infatti, quasi tutti sono disposti a scommettere in un gioco equo 5 Euro per vincere 5000 Euro, ma pochi investirebbero 5000 Euro per vincere 5 milioni di Euro; eppure, è sempre pari a 0.001 la valutazione della probabilità misurata come rapporto tra posta e vincita.

Escludendo la definizione classica (importante ma circoscritta), va aggiunto che le impostazioni frequentista e soggettivista della Teoria della probabilità sono abbastanza contrapposte perché derivano da scuole filosofiche diverse:

- per i *frequentisti* la probabilità è un requisito intrinseco degli eventi e degli esperimenti che li generano (come l'altezza di un albero o il peso di una pietra);
- per i *soggettivisti* la probabilità è una libera interpretazione dell'Uomo in rapporto agli eventi.

In tale ottica, esse sembrano inconciliabili e, in talune circostanze, producono differenti conseguenze statistiche, soprattutto nel momento inferenziale. Tuttavia, ciò che conta ai nostri fini è che tutte le impostazioni conducano – quasi esattamente – alla medesima assiomatizzazione del Calcolo delle probabilità e, di conseguenza, alla stessa misura della probabilità nelle applicazioni più comuni.

8.5 Assiomatizzazione del Calcolo delle probabilità

Una scienza assiomatica stabilisce il proprio contenuto individuando gli elementi di cui si occupa, le regole per combinarli e gli enunciati fondamentali, ricercando poi nuovi risultati (i teoremi) per ampliare il proprio campo di interesse, spesso sulla spinta di esigenze applicative.

o Una scienza assiomatica non è, di per sé, agganciata alla realtà fenomenica nel senso che gli elementi di cui si occupa non sono necessariamente oggetti o proposizioni riguardanti il mondo fisico; si pensi, per esempio, all'Algebra moderna o alla Logica. Tuttavia, essa diventa utile allorquando fornisce soluzioni e proposte per l'analisi del mondo reale, anche per analogia o mediante approssimazioni. Peraltro, anche se le presentazioni

scientifiche di tipo assiomatico costituiscono sistemi eleganti e coerenti, esse sono spesso il risultato di un lungo travaglio intellettuale generato dall'osservazione. Per esempio, le proprietà dei triangoli sono formali e rigorosamente dimostrabili nell'ambito della Geometria euclidea, ma questi risultati sono il frutto della osservazione condotta su figure reali e sono stati derivati da problemi molto concreti.

Questi principi erano ben evidenti a Kolmogorov che, all'inizio della sua opera fondamentale («*The Foundations of Probability*», 1933), così scriveva:

The theory of probability, as a mathematical discipline, can and should be developed from axioms in exactly the same way as Geometry and Algebra. This means that after we have defined the elements to be studied and their basic relations, and have stated the axioms by which these relations are to be governed, all further exposition must be based exclusively on these axioms, independent of the usual concrete meaning of these elements and their relations.

I momenti fondamentali di ogni assiomatizzazione sono i seguenti:

1. *Individuazione dei* ***concetti primitivi***, cioè di quegli elementi non definibili ma per i quali la storia, la biologia, la cultura e l'esperienza degli esseri umani hanno prodotto convinzioni universali sì da farli ritenere coincidenti.

   o Per l'Aritmetica, il numero è un concetto primitivo, così come per la Geometria lo sono il punto, la retta, il piano, lo spazio, e per la Fisica lo è il corpo. I concetti primitivi sono elementi del linguaggio corrente a cui la scienza fornisce un significato più esplicito solo a fini di formalizzazione matematica. Quindi, i concetti primitivi non si definiscono ma al più si esplicitano e si circoscrivono, anche con esempi, per chiarire eventuali ambiguità linguistiche.

2. *Enunciazione dei* ***postulati*** (assiomi), cioè di affermazioni non dimostrabili ma utili e coerenti per costruire la scienza assiomatica. Ciò che rende un'affermazione di interesse scientifico un postulato è il fatto che la sua dimostrazione necessita del postulato stesso, per cui esso può essere solo assunto come un «principio ineliminabile».

   o In Aritmetica, è un assioma il postulato archimedeo secondo cui l'aggiunta di un'unità a ciascun numero produce ancora un numero; in Geometria, è famoso il postulato delle parallele; in Fisica, è un postulato quello della incompenetrabilità dei corpi solidi. I postulati sono affermazioni dogmatiche la cui validità non è derivabile da altre affermazioni e costituiscono la base per ogni sviluppo successivo.

3. *Dimostrazione dei* ***teoremi*** mediante i postulati e, ovviamente, con l'ausilio della Logica e della Matematica.

   o In Aritmetica, sono teoremi le affermazioni: «il quadrato di un numero che termina con 5 termina sempre con 25» oppure che: «la somma dei primi n numeri dispari è uguale al quadrato di n». Similmente, per la Geometria, sono dei teoremi le seguenti affermazioni: «la somma degli angoli di un triangolo è 180°» oppure «la somma dei quadrati costruiti sui cateti di un triangolo rettangolo è uguale al quadrato costruito sull'ipotenusa». In Fisica, invece, i teoremi sono le cosiddette «leggi» come la relazione tra spazio, tempo e velocità di un corpo che cade non soggetto ad altre forze.

Nel seguito, espliciteremo le varie fasi necessarie per la costruzione assiomatica del Calcolo delle probabilità, che avviene in forma compiuta grazie al lavoro del 1933 di Kolmogorov e che è stata riproposta e diffusa in Italia da Pompilj, assieme a rilievi critici e ampliamenti concettuali.

In tale ambito, il prossimo paragrafo tratta dei concetti primitivi mentre i paragrafi successivi saranno dedicati alla formalizzazione delle prove e degli eventi, rispettivamente. Quindi, enunceremo i postulati (paragrafo 8.9) e, infine, presenteremo i principali teoremi del Calcolo delle probabilità (paragrafo 8.10).

8.6 Concetti primitivi

I concetti primitivi del Calcolo delle probabilità sono ***prova***, ***evento***, ***probabilità*** e sono sequenzialmente richiamati con grande efficacia nella frase: *la prova genera l'evento con una certa probabilità* (Pompilj).

I concetti primitivi sono innati per cui quelle che seguono non sono definizioni ma solo esempi e chiarimenti per consentire, con una certa precisione terminologica, di riferirsi esattamente allo stesso concetto usando quelle parole.

■ Per ***prova*** si intende ogni esperimento soggetto ad incertezza. Spesso, per prova si intende il meccanismo fisico che produce i risultati, per cui la conoscenza della prova equivale alla conoscenza delle circostanze che la determinano. Nei casi più complessi, una prova può essere articolata in più fasi che si definiscono *sottoprove*, ciascuna delle quali deve essere ben definita in quanto prova. In senso lato, prova è anche ogni affermazione di cui non è possibile decidere univocamente della verità o falsità, per mancanza di elementi conoscitivi essenziali.

o Costituiscono esempi di prove di interesse per il Calcolo delle probabilità le seguenti: il lancio di un dado; l'estrazione di una carta da un mazzo di carte; il presentarsi all'esame di Statistica in una certa sessione; gli esperimenti di laboratorio per accertare se il farmaco A è preferibile al farmaco B; e così via. In tutti questi casi le modalità dell'esperimento rendono incerti gli esiti possibili. In senso lato, costituisce una prova anche l'attribuzione di un testo anonimo ad un certo Autore scelto in una lista.

o Sono prove composte da più sottoprove le seguenti: il lancio di due dadi; il lancio di una moneta sino a che esce Testa per la prima volta; l'estrazione di due palline da un'urna; la scelta di 200 persone dagli elenchi telefonici delle 10 maggiori città italiane, e così via. Tuttavia, mentre nei primi due esempi la dizione rende chiari ed univoci gli esperimenti, nei successivi essi vanno accuratamente definiti per una corretta utilizzazione probabilistica.

o Più precisamente, l'estrazione di due palline da un'urna non costituisce una prova per il Calcolo delle probabilità sino a che non si precisa se, nella seconda estrazione, la pallina già estratta viene rimessa nell'urna, oppure no. Similmente, la scelta delle 200 persone dai 10 elenchi non costituisce una prova sino a che non si chiarisce la ripartizione tra le 10 città delle 200 persone da estrarre (in parti uguali, in proporzione ai residenti, etc.) e, poi, in quale modo da ciascun elenco si procederà alla individuazione delle persone da estrarre (1 persona ogni 100 pagine, tutte le persone da una sola pagina scelta a caso, etc.).

■ Per ***evento*** si intende uno dei possibili risultati della prova ed esso consiste nella descrizione precisa ed esauriente delle sue caratteristiche in rapporto alla natura e allo svolgimento dell'esperimento. Poiché le prove di cui si interessa il Calcolo delle probabilità sono quelle soggette ad incertezza, esiste sempre una pluralità di eventi come possibili risultati di una prova. Da questo punto di vista, ne consegue che una prova è ben definita se e solo se sono ben definiti tutti gli eventi cui essa può dar luogo. Pertanto, anche se nel linguaggio corrente esiste una tendenza alla semplificazione del discorso, non è corretto parlare di eventi, in astratto, senza riferirsi ad una specifica prova.

o Nella lingua italiana, «evento» indica un accadimento cui si attribuisce un significato importante (evento storico, lieto evento, etc.) mentre il senso del concetto primitivo di cui discutiamo è semplicemente quello di «risultato di un esperimento». Nella terminologia anglosassone, si distingue tra «*event*» nel primo senso e «*outcome*» nel senso del Calcolo delle probabilità. Il termine «outcome» (venire fuori, uscire) è particolarmente indicato per chiarire la natura dell'evento in rapporto alla prova; infatti, anche noi diciamo, per esempio: «sulla ruota di Firenze è uscito il 27» per esplicitare il presentarsi di un evento particolare nella prova «estrazione del Lotto sulla ruota di Firenze».

o Nella prova «lancio di una moneta», l'evento «Testa» è uno dei 2 possibili eventi; l'altro è l'evento «Croce».

o Nella prova «lancio di due monete» gli eventi possibili sono quattro: «Testa sulla prima moneta e Croce sulla seconda», «Testa sulla prima e Testa sulla seconda», «Croce sulla prima e Testa sulla seconda», «Croce sulla prima e Croce sulla seconda». In questa stessa prova, sono anche eventi: «si verifica almeno una Testa», «si verifica solo una Testa», «si verificano due facce uguali», e così via.

o Nella prova «sostenere l'esame di Statistica e ricevere la relativa valutazione» sono eventi «l'esame è superato», «ricevo un voto di 28/30», «ricevo un voto non inferiore a 25/30», e così via.

o Nella prova «lancio di due dadi» sono eventi: «si verifica un doppio 6», «la differenza in valore assoluto tra i punti sulle due facce è superiore a 3», «si verifica un numero differente di punti sulle due facce», e così via.

Come si vede, talvolta l'evento viene descritto in modo semplice mentre in altri casi appare più complesso (più oltre, distingueremo tra *evento elementare* ed *evento composto*). In ogni caso, ogni evento va riferito ad una prova ben definita. Per esempio, l'evento «una Testa» nella prova «lancio di una moneta» è completamente diverso dall'evento «una Testa» nella prova «lancio di due monete».

■ Per ***probabilità*** si intende un numero associato al verificarsi di un evento. Tale affermazione non definisce la probabilità (perché è un concetto primitivo) ma circoscrive l'uso che ne faremo, evidenziando in che senso essa sia associata al risultato di una prova.

o Tra i concetti primitivi che costellano la nostra esperienza (spazio, numero, tempo, etc.), la probabilità è uno dei più complessi e più lenti ad emergere sia nella evoluzione personale del bambino che nella storia collettiva della Umanità perché l'incertezza genera paure ed insicurezze e, quindi, nello sviluppo iniziale della conoscenza essa viene rimossa ed accuratamente evitata. È solo nella fase matura della evoluzione, sia dell'individuo che della specie umana, che ci si rende conto della possibilità di controllare anche l'incertezza mediante il Calcolo delle probabilità, accrescendo alla fine la nostra conoscenza del mondo esterno.

Si osservi che, pur essendo un concetto primitivo (e, quindi, vago perché non definibile), la probabilità è certamente un numero perché chi pensa alla probabilità istituisce sempre – anche senza saperlo – una relazione d'ordine (maggiore, minore, uguale) tra le probabilità di eventi differenti, collocando la probabilità di un evento specifico tra quella di due «eventi estremi»: un evento che non può mai verificarsi ed un evento che si verifica sempre.

8.7 Formalizzazione delle prove

Ogni prova consiste nella descrizione di una serie di azioni da compiere o di fatti da osservare che, in modo univoco, consentono la elencazione esaustiva dei possibili eventi che essa può generare. Una prova è completa quando si è svolta nelle circostanze spazio-temporali che la definiscono.

Se una prova $\mathcal{P}$ consiste di due sottoprove $\mathcal{P}_1$ e $\mathcal{P}_2$ scriveremo: $\mathcal{P} = \mathcal{P}_1 \times \mathcal{P}_2$ e diremo che la prova si è completata quando sia $\mathcal{P}_1$ che $\mathcal{P}_2$ sono state ordinatamente completate. *Questa affermazione*, a prima vista banale, *costituisce un punto importante nella formalizzazione del Calcolo delle probabilità.*

La simbologia $\mathcal{P} = \mathcal{P}_1 \times \mathcal{P}_2$ esprime $\mathcal{P}$ come il prodotto cartesiano di $\mathcal{P}_1$ e $\mathcal{P}_2$ e indica che la prova assumerà come risultati possibili tutte le coppie ordinate di eventi (E_{1i}, E_{2j}), ove E_{1i} è generato dalla sottoprova $\mathcal{P}_1$ ed E_{2j} è generato dalla sottoprova $\mathcal{P}_2$.

o Se la prova $\mathcal{P} = \mathcal{P}_1 \times \mathcal{P}_2$ consiste nell'estrazione di due palline da un'urna, che contiene palline Bianche e Nere, e si è interessati all'evento «la seconda pallina è Bianca», non sarà possibile specificare il risultato della sottoprova $\mathcal{P}_2$, relativa alla seconda pallina, senza definire *anche* il risultato relativo alla sottoprova $\mathcal{P}_1$, relativa alla prima pallina estratta. Infatti, se una prova $\mathcal{P}$ consiste nell'estrazione di due palline, l'evento E: «la seconda pallina è Bianca» è un risultato di $\mathcal{P}$ se e solo se esso è esprimibile come risultato della prova $\mathcal{P} = \mathcal{P}_1 \times \mathcal{P}_2$. Quindi, l'evento «la seconda pallina è Bianca» va correttamente espresso nel modo seguente: «(Bianca oppure Nera) nella prima estrazione e (Bianca) nella seconda estrazione».

È praticamente impossibile formalizzare tutte le prove che si possono immaginare, per cui ci si limita a descrivere alcune categorie di esperimenti che hanno il vantaggio di includere un gran numero di situazioni reali che, a prima vista, appaiono molto differenti tra loro.

• Lo schema di ***estrazione da un'urna di una o più palline***, tutte uguali eccetto per qualche caratteristica (numero o colore, per lo più), è una prova che ha il pregio della semplicità concettuale ma che, nelle numerose varianti, è malleabile al punto da essere una adeguata formalizzazione per la quasi totalità dei giochi più comuni, per i sondaggi demoscopici, per gli esperimenti da laboratorio e per

numerose altre situazioni ove è richiesta una designazione casuale dei soggetti.

Se la prova si distingue in più sottoprove, è fondamentale distinguere le modalità di estrazione consecutiva delle palline.

i) Quando in ciascuna sottoprova l'urna presenta la stessa composizione si parla di *estrazioni bernoulliane* (o «con ripetizione», perché in ciascuna sottoprova una pallina precedentemente estratta può essere ri-estratta e quindi «ripetersi»). In questa tipologia di estrazione, la struttura probabilistica dell'esperimento rimane identica in tutte le sottoprove.

ii) Quando in ciascuna sottoprova la pallina estratta non viene rimessa nell'urna si parla di *estrazioni in blocco* (o «senza ripetizione», perché in ciascuna sottoprova la pallina estratta «non può ripetersi» essendo esclusa dalle sottoprove successive). In questa tipologia di estrazione, la struttura probabilistica dell'esperimento si modifica ad ogni sottoprova.

- Il lancio di una moneta ben equilibrata è assimilabile ad una estrazione da un'urna che contiene palline Bianche e Nere in pari proporzione.
- La prova «lancio di un dado ben equilibrato» equivale all'estrazione di una pallina da un'urna che contiene sei palline identiche ma numerate da 1 a 6.
- L'intervista a dei cittadini per chiedere loro se sono favorevoli o contrari ad un referendum abrogativo equivale ad estrarre un sottoinsieme di palline da un'urna in cui è incognita la proporzione di palline Bianche (i favorevoli) e Nere (i contrari).
- I quattro giochi seguenti: «lancio di una moneta», «lancio di un dado», «scelta di una carta da un mazzo di carte napoletane», «estrazione di un numero al Lotto» sono tutti formalizzabili nella prova: «estrazione di una pallina da un'urna che contiene palline numerate da 1 a n», dove $n = 2$ per la moneta, $n = 6$ per il dado, $n = 40$ per il mazzo di carte napoletane, $n = 90$ per il Lotto. Quindi, se si studia la prova consistente nell'estrazione di una pallina da un'urna che ne contiene n si introduce uno schema generale che può essere poi specificato a seconda dell'esperimento che interessa.
- Se si estraggono 2 palline da un'urna che contiene 1 pallina Bianca e 9 Nere, occorre chiarire le modalità di estrazione della seconda pallina: altrimenti, la prova non è correttamente specificata. Di fatto, le modalità di estrazione determinano l'insieme dei possibili eventi che si possono verificare e, quindi, le corrispondenti probabilità. Così, se l'estrazione è bernoulliana (cioè si rimette nell'urna la pallina estratta prima di estrarre la seconda, limitandosi a registrarne il colore), allora l'evento «le due palline estratte sono Bianche» può effettivamente verificarsi, anche se con una probabilità alquanto piccola. Invece, se l'estrazione è in blocco (cioè, la prima pallina estratta non viene rimessa nell'urna e si estrae la seconda dall'urna che ora ne contiene 9: ciò equivale ad estrarre contemporaneamente un solo blocco di 2 palline) l'evento «le due palline estratte sono Bianche» non può mai verificarsi perché nell'urna vi è 1 sola pallina Bianca. Quindi, nell'estrazione in blocco l'evento «due palline Bianche» non è un risultato ammissibile per tale prova.

- Un'altra tipologia di prove è quella dell'***attesa dell'evento***, generalmente secondo due specificazioni tra loro collegate: in un caso, trattasi di prove ripetute «sino a che l'evento non si verifica»; nell'altro, trattasi «dell'intervallo di tempo intercorrente fra il presentarsi di due eventi».

- Nella prova «lanciare una moneta sino a che esca Testa per la prima volta», le sottoprove che formano il risultato di tale prova sono «primo lancio» (se esce immediatamente Testa), «secondo lancio» (se esce prima Croce e poi Testa), «terzo lancio» (se escono due Croce e poi Testa), ..., «m-esimo lancio» (se per $m-1$ volte esce Croce e poi esce Testa), ..., e così via all'infinito. Poiché i possibili risultati della prova sono in corrispondenza con la serie naturale $(1, 2, 3, \ldots, m, \ldots)$ questa prova genera una infinità *numerabile* di eventi.
- La prova consistente nell'esaminare l'intervallo di tempo tra una telefonata e la successiva per un certo abbonato consiste in un esperimento che genera gli eventi «la nuova telefonata è giunta dopo un tempo t dalla precedente», per ogni $t \in \mathbb{R}^+$, essendo $\mathbb{R}^+$ l'insieme dei numeri reali positivi. In questo caso, i possibili risultati della prova sono in corrispondenza con un intervallo reale e quindi costituiscono una infinità continua (*non-numerabile*) di eventi.

- Una ulteriore caratterizzazione delle prove deriva dai cosiddetti problemi di ***probabilità geometrica*** nei quali gli eventi sono costituiti da enti geometrici (definiti sulla retta, sul piano, nello spazio e, in generale, in $\mathbb{R}^n$) e la prova consiste nell'esaminare la possibilità che si verifichi una relazione tra enti geometrici. L'utilità di questi problemi si manifesta in Botanica, Ecologia, Fisica, etc.

- Nella prova «lancio di una freccia su un bersaglio circolare», supponendo che certamente l'arciere colpisca il bersaglio, si può esaminare l'evento «la freccia colpisce un punto più vicino al centro che alla circonferenza che delimita il bersaglio» oppure l'evento «la freccia non colpisce alcun punto del quadrato inscritto nel bersaglio circolare», e così via.
- Nella prova «scelta a caso di due punti su un bastoncino di legno di lunghezza L» si considera l'evento «con i tre segmenti risultanti è possibile costruire un triangolo» oppure «con i tre segmenti il triangolo che eventualmente si forma avrà area superiore a $L^2/100$», e così via.

Tali problemi astratti hanno applicazioni reali non trascurabili, tra le quali ci limitiamo a segnalare le seguenti:

i) la probabilità di attecchimento e di sviluppo di un seme trasportato dal vento dipende dalla probabilità che esso non ricada in un cerchio che si sovrappone con il cerchio minimo vitale per una specie in competizione;

ii) l'efficacia di un antiparassitario somministrato in punti specifici dipende dalle dimensioni del suo raggio di azione;

iii) la visibilità di un animale selvaggio in una foresta si può tradurre in un problema di probabilità geometrica.

8.8 Formalizzazione degli eventi

La formalizzazione degli eventi potrebbe essere risolta mediante il seguente postulato: «*Gli eventi formano un'algebra di Boole completa*».

o Infatti, esso costituisce un'affermazione inequivoca per chi già conosce l'analisi reale ma, per motivi didattici, è opportuna una breve illustrazione di tale struttura (l'algebra di Boole completa) orientata verso il Calcolo delle probabilità.

Nel seguito, introdurremo gli elementi dell'algebra (cioè gli eventi) e, quindi, le operazioni tra tali elementi (unione, intersezione e negazione), anche mediante esempi e rappresentazioni grafiche.

In generale, un'*algebra* è un insieme di enti, di operazioni e di regole per collegare gli elementi della stessa algebra. L'algebra degli eventi è, quindi, una struttura formale per collegare gli eventi tra di loro.

Gli eventi si indicano con le prime lettere maiuscole dell'alfabeto latino ($A, B, C, D, E, \ldots$), talvolta con un indice quando vi è una successione: $E_1, E_2, \ldots$, oppure: E_i, $i = 1, 2, \ldots$

Per ragioni didattiche, si introducono tre operazioni (*unione, intersezione* e *negazione*), ma le operazioni *essenziali* per definire un'algebra degli eventi sono solo due (negazione e unione, oppure negazione e intersezione).

■ Definiamo ***unione*** tra due eventi A e B quell'evento C che si verifica allorquando si verifica *almeno* uno dei due eventi A e B: cioè se si verifica A, se si verifica B o se si verificano contemporaneamente A e B (quando ciò è possibile).

L'unione tra A e B si indica con: $A \cup B$ (che si legge: «A o B» oppure: «A unito B»).

L'unione si può estendere ad un numero finito o numerabile di eventi; così la scrittura: $A \cup B \cup C$ indica quell'evento che si verifica quando si verifica almeno un evento tra A, B, C, cioè se si verifica A oppure B oppure C oppure qualsiasi combinazione dei tre eventi a due a due oppure a tre a tre (se ciò è possibile).

Data una collezione di eventi $E_1, E_2, \ldots$ indicheremo con: $E_1 \cup E_2 \cup \ldots$ l'unione tra tutti gli eventi della collezione, cioè «almeno uno tra gli eventi $E_1, E_2, \ldots$».

Se la collezione consiste di un insieme numerabile potremo anche scrivere:

$$E_1 \cup E_2 \cup \cdots \cup E_n \cup \ldots = \bigcup_{i=1}^{\infty} E_i$$

per indicare che, data la successione di eventi: $E_i, i = 1, 2, \ldots$, si verifica $\bigcup_{i=1}^{\infty} E_i$ se si verifica almeno uno tra gli eventi E_i, $i = 1, 2, \ldots$

Esempio 8.1

La prova «lancio di un dado» genera i 6 possibili eventi: «faccia con 1 punto», «faccia con 2 punti», ..., «faccia con 6 punti» che indicheremo con: $E_1, E_2, \ldots, E_6$, rispettivamente.

Definiamo, ora, gli eventi seguenti:

$$A = \{E_1, E_2\};\ B = \{E_3, E_4, E_5\};\ C = \{E_1, E_3, E_5\};\ D = \{E_2, E_4, E_6\}.$$

Sono, allora, immediati i seguenti risultati:

$$\begin{aligned} A \cup B &= \{E_1, E_2\} \cup \{E_3, E_4, E_5\} = \{E_1, E_2, E_3, E_4, E_5\}; \\ A \cup C &= \{E_1, E_2\} \cup \{E_1, E_3, E_5\} = \{E_1, E_2, E_3, E_5\}; \\ B \cup C &= \{E_3, E_4, E_5\} \cup \{E_1, E_3, E_5\} = \{E_1, E_3, E_4, E_5\}; \\ A \cup B \cup D &= \{E_1, E_2\} \cup \{E_3, E_4, E_5\} \cup \{E_2, E_4, E_6\} = \{E_1, E_2, E_3, E_4, E_5, E_6\}. \end{aligned}$$

□

È importante distinguere i risultati di una prova in *eventi elementari* ed *eventi composti*. La definizione più semplice è per esclusione: *un evento elementare è quello che non può mai essere scritto come unione di altri eventi elementari*; ne consegue che *un evento composto può essere scritto come unione di più eventi elementari*.

Inoltre, occorre puntualizzare cosa si intende esattamente con la frase: *si è verificato l'evento E*. D'ora innanzi, l'espressione «*E* ***si è verificato***» significa: ***si è verificato almeno un evento elementare di cui E costituisce l'unione.*** Questa precisazione è importante per il calcolo corretto delle probabilità di eventi composti.

Esempio 8.2

Nella prova «lancio di un dado», esistono 6 eventi elementari che indichiamo con $E_1, E_2, E_3, E_4, E_5, E_6$. Siano, ora, A e B i due seguenti eventi:

$$\begin{aligned} A &= \{\text{il dado mostra la faccia con 2 punti}\}; \\ B &= \{\text{il dado mostra una faccia con un numero pari di punti}\}. \end{aligned}$$

Ora, l'espressione «l'evento A si è verificato» è ben evidente perché indica che, tra i 6 possibili risultati della prova, si è realizzato quell'unico evento elementare E_2 di cui A è costituito. Invece, nel secondo caso, l'affermazione «B si è verificato» indica che una sola delle 3 facce cui B fa riferimento si è presentata perché: $B = E_2 \cup E_4 \cup E_6$. Quindi, l'affermazione «B si è verificato» coincide con l'affermazione «almeno uno degli eventi elementari di B si è verificato», cioè si è verificata una delle tre facce con un numero pari di punti.

□

■ Definiamo ***negazione*** dell'evento A quell'evento che si verifica quando non si verifica A, e lo indicheremo con $\overline{A}$ (che si legge: «non A» oppure «A negato»).

È evidente che la negazione della negazione di un qualsiasi evento A coincide con l'evento A, cioè: $\overline{\overline{A}} = A$.

o Si osservi che mentre la unione tra A e B è una operazione che fà riferimento ai due soli eventi A e B (senza riferirsi agli altri eventi di cui la prova consiste), la negazione di un evento A fa riferimento a tutti gli eventi della prova eccetto l'evento A.

Indichiamo con Ω lo ***spazio campione***, cioè l'insieme di tutti i risultati di una prova. Allora, Ω è un evento che si verifica sempre, perché in ogni caso si verifica uno degli eventi elementari di cui si compone. Quindi, Ω è l'***evento certo.***

La negazione di Ω non si verificherà mai, per cui $\overline{\Omega}$ è l'***evento impossibile*** e sarà indicato con $\emptyset$. Ovviamente, la negazione di $\emptyset$ sarà l'evento certo: $\overline{\emptyset} = \Omega$.

A differenza di tutti gli altri eventi (il cui significato deriva esclusivamente dalla prova che li genera), gli eventi Ω e $\emptyset$ sono gli unici eventi ben definiti a prescindere dalla struttura dell'esperimento probabilistico di cui si discute.

Esempio 8.3

Nella prova «lancio di due dadi» si generano 36 possibili risultati espressi dalle possibili disposizioni di ciascuna delle 6 facce del 1° dado con le corrispondenti 6 facce del 2° dado.

Supponiamo, in tale prova, che si scommetta sulla somma dei punti che si verificano sulle due facce risultanti. Evidentemente, la somma assumerà qualsiasi valore intero compreso tra 2 e 12, per cui indichiamo con S_j l'evento «la somma dei punti è il numero j», per $j = 2, 3, \ldots, 12$.

Allora, l'evento certo è:

$$\Omega = \{S_2, S_3, S_4, S_5, S_6, S_7, S_8, S_9, S_{10}, S_{11}, S_{12}\}$$

perché costituisce la collezione di tutti i possibili risultati di tale prova. Ovviamente, esso è «certo» perché si verifica sempre, cioè si verifica sempre uno degli eventi elementari di cui Ω è composto (chi scommette su Ω vince sempre!).

Su tale prova, definiamo gli eventi seguenti:

$$\begin{aligned}
A &= \{\textit{la somma dei punti è al massimo pari a 4}\} = \{S_2, S_3, S_4\};\\
B &= \{\textit{la somma dei punti è inferiore a 9}\} = \{S_2, S_3, S_4, S_5, S_6, S_7, S_8\}.\\
C &= \{\textit{la somma dei punti è superiore a 4}\} = \{S_5, S_6, S_7, S_8, S_9, S_{10}, S_{11}, S_{12}\}.\\
D &= \{\textit{la somma dei punti è 6 oppure 7}\} = \{S_6, S_7\}.
\end{aligned}$$

Sono, allora, immediatamente derivabili le seguenti unioni tra eventi:

$$\begin{aligned}
A \cup B &= \{S_2, S_3, S_4, S_5, S_6, S_7, S_8\};\\
A \cup C &= \{S_2, S_3, S_4, S_5, S_6, S_7, S_8, S_9, S_{10}, S_{11}, S_{12}\} = \Omega;\\
B \cup C &= \{S_2, S_3, S_4, S_5, S_6, S_7, S_8, S_9, S_{10}, S_{11}, S_{12}\} = \Omega;\\
A \cup B \cup D &= \{S_2, S_3, S_4, S_5, S_6, S_7, S_8\}.
\end{aligned}$$

Similmente, si deducono le seguenti negazioni di eventi:

$$\begin{aligned}
\overline{A} &= \{\text{tutti gli eventi di } \Omega \text{ eccetto } S_2, S_3, S_4\} = \{S_5, S_6, S_7, S_8, S_9, S_{10}, S_{11}, S_{12}\};\\
\overline{B} &= \{\text{la somma dei punti è non inferiore a 9}\} = \{S_9, S_{10}, S_{11}, S_{12}\};\\
\overline{C} &= \{\text{la somma dei punti è non superiore a 4}\} = \{S_2, S_3, S_4\};\\
\overline{D} &= \{\text{la somma dei punti è diversa da 6 e 7}\} = \{S_2, S_3, S_4, S_5, S_8, S_9, S_{10}, S_{11}, S_{12}\}.
\end{aligned}$$

Naturalmente, sono anche possibili unioni e negazioni combinate tra loro in vario modo, per le quali si suggerisce di sviluppare ordinatamente le operazioni tra eventi, utilizzando le parentesi quando vi sono dubbi. Per esempio, si ottengono i seguenti risultati:

$$\begin{aligned}
\overline{B} \cup \overline{C} &= \{S_9, S_{10}, S_{11}, S_{12}\} \cup \{S_2, S_3, S_4\} = \{S_2, S_3, S_4, S_9, S_{10}, S_{11}, S_{12}\};\\
\overline{A} \cup \overline{B} \cup \overline{D} &= \{S_2, S_3, S_4, S_5, S_6, S_7, S_8, S_9, S_{10}, S_{11}, S_{12}\} = \Omega;\\
\overline{A \cup B} &= \overline{\{S_2, S_3, S_4, S_5, S_6, S_7, S_8\}} = \{S_9, S_{10}, S_{11}, S_{12}\};\\
\overline{\overline{B} \cup \overline{C}} &= \overline{\{S_2, S_3, S_4, S_9, S_{10}, S_{11}, S_{12}\}} = \{S_5, S_6, S_7, S_8\};\\
\overline{B \cup C} &= \overline{\Omega} = \emptyset\,.
\end{aligned}$$

□

• Una collezione $\mathcal{A}$ di eventi $E_1, E_2, \ldots, E_n, \ldots$ è un'***algebra di Boole completa*** se valgono le due seguenti condizioni:

i) se $E_i \in \mathcal{A}, i = 1, 2, \ldots$, allora $\overline{E_i} \in \mathcal{A}$;

ii) se $E_i \in \mathcal{A}, i = 1, 2, \ldots$, allora $\bigcup_{i=1}^{n} E_i \in \mathcal{A}, \forall n = 1, 2, \ldots$

Quindi, la collezione $\mathcal{A}$ di eventi è un'algebra di Boole *completa* se essa è chiusa rispetto alla negazione e all'unione di una quantità numerabile di eventi (se ciò avviene solo per n finito, essa si chiama semplicemente «algebra di Boole»). Un'algebra di Boole completa si definisce anche *σ-algebra.*

■ Definiamo ***intersezione*** di A e B quell'evento C che si verifica se e solo se si verificano *contemporaneamente* sia A che B, cioè se A e B assieme costituiscono il risultato della prova. Scriveremo: $C = A \cap B$ (che si legge: «A e B» oppure: «A intersecato con B»).

Esempio 8.4

Nella prova descritta nell'esempio 8.1, ricordando gli eventi A, B, C, D ivi definiti, si ottiene subito:

$$\begin{aligned}
A \cap B &= \{E_1, E_2\} \cap \{E_3, E_4, E_5\} = \{\text{nessun evento della prova}\} = \emptyset;\\
A \cap C &= \{E_1, E_2\} \cap \{E_1, E_3, E_5\} = \{E_1\};\\
A \cap C \cap D &= (\{E_1, E_2\} \cap \{E_1, E_3, E_5\}) \cap \{E_2, E_4, E_6\} = \{E_1\} \cap \{E_2, E_4, E_6\} = \emptyset;\\
B \cap C &= \{E_3, E_4, E_5\} \cap \{E_1, E_3, E_5\} = \{E_3, E_5\}.
\end{aligned}$$

□

Esempio 8.5

Riprendendo l'esempio 8.3, e gli eventi A, B, C, D ivi definiti, è agevole dedurre che:

$$\begin{aligned} A \cap B &= \{S_2, S_3, S_4\} \cap \{S_2, S_3, S_4, S_5, S_6, S_7, S_8\} = \{S_2, S_3, S_4\}; \\ A \cap C &= \{S_2, S_3, S_4\} \cap \{S_5, S_6, S_7, S_8, S_9, S_{10}, S_{11}, S_{12}\} = \emptyset\ ; \\ B \cap C &= \{S_2, S_3, S_4, S_5, S_6, S_7, S_8\} \cap \{S_5, S_6, S_7, S_8, S_9, S_{10}, S_{11}, S_{12}\} = \{S_5, S_6, S_7, S_8\} \end{aligned}$$

Si possono, inoltre, combinare opportunamente le varie operazioni sin qui introdotte. Per esempio:

$$\begin{aligned} (A \cap B) \cup D &= \{S_2, S_3, S_4\} \cup \{S_6, S_7\} = \{S_2, S_3, S_4, S_6, S_7\}; \\ \overline{A} \cap B &= \overline{\{S_2, S_3, S_4\}} \cap \{S_2, S_3, S_4, S_5, S_6, S_7, S_8\} = \{S_5, S_6, S_7, S_8\}; \\ A \cap \overline{B} &= \{S_2, S_3, S_4\} \cap \overline{\{S_2, S_3, S_4, S_5, S_6, S_7, S_8\}} = \emptyset\ ; \\ \left(\overline{A} \cap B\right) \cup \left(A \cap \overline{B}\right) &= \{S_5, S_6, S_7, S_8\} \cup \emptyset = \{S_5, S_6, S_7, S_8\}; \\ \overline{A} \cap \overline{B} &= \overline{\{S_2, S_3, S_4\}} \cap \overline{\{S_2, S_3, S_4, S_5, S_6, S_7, S_8\}} = \{S_9, S_{10}, S_{11}, S_{12}\}; \\ \overline{\overline{A} \cap \overline{B}} &= \overline{\{S_9, S_{10}, S_{11}, S_{12}\}} = \{S_2, S_3, S_4, S_5, S_6, S_7, S_8\}. \end{aligned}$$

□

Il legame fra le tre operazioni è offerto dalle seguenti relazioni dette *leggi di de Morgan*:

$$\overline{A \cap B} = \overline{A} \cup \overline{B}; \qquad \overline{A \cup B} = \overline{A} \cap \overline{B}.$$

∘ Esse sono due relazioni di cui una è *duale* dell'altra, perché ciascuna si ottiene dall'altra scambiando le operazioni di unione e di intersezione, rispettivamente. Si leggono: «la negazione della intersezione tra due eventi è pari alla unione delle rispettive negazioni» e, similmente, «la negazione della unione tra due eventi è pari alla intersezione delle rispettive negazioni».

∘ Le leggi di de Morgan si possono generalizzare ad un numero qualsiasi, anche infinito numerabile, di unioni ed intersezioni. Tutte le volte che gli eventi $E_i, i = 1, 2, \ldots$ sono ben definiti per una specifica prova, valgono infatti le seguenti formule generali:

$$\overline{\bigcap E_i} = \bigcup \overline{E_i}; \qquad \overline{\bigcup E_i} = \bigcap \overline{E_i}.$$

∘ Le leggi di de Morgan sono relazioni onnipresenti nelle scienze formali (Aritmetica, Geometria, Analisi, Probabilità, etc.) e si dimostrano secondo i principi generali della Logica, che qui omettiamo. Si può intuire, però, nel caso di due eventi la ragionevolezza delle relazioni proposte. L'affermazione: «non è vero che si verifica A contemporaneamente a B» coincide con l'affermazione: «non si verifica A oppure non si verifica B»; e similmente per la seconda legge.

∘ La ragionevolezza delle leggi del de Morgan emerge con maggiore evidenza se si fa riferimento a qualche situazione concreta. Per esempio, nella prova «il tempo che farà domani», definiamo gli eventi $A = \{piove\}$ e $B = \{tira\ vento\}$; allora, è agevole controllare che l'evento:

$\overline{A \cap B} = \overline{\{piove\ e\ tira\ vento\}} = \{non\ avverrà\ che,\ contemporaneamente,\ piova\ e\ tiri\ vento\}$

coincide con l'evento:

$$\overline{A} \cup \overline{B} = \overline{\{piove\}} \cup \overline{\{tira\ vento\}} = \{avverrà\ che\ o\ non\ piova\ o\ non\ tiri\ vento\}.$$

∘ Le leggi di de Morgan confermano che sono solo due le operazioni essenziali dell'algebra degli eventi; infatti, negando ambo i membri delle due leggi, si ottiene:

$$A \cap B = \overline{\overline{A} \cup \overline{B}}; \qquad A \cup B = \overline{\overline{A} \cap \overline{B}}.$$

Quindi, la prima relazione consente di definire l'intersezione se sono già definite negazione e unione, mentre la seconda consente di definire l'unione se sono già definite negazione e intersezione. Da ciò discende che l'algebra di Boole completa è chiusa anche rispetto all'intersezione, finita o numerabile, di una collezione di eventi perché è sempre possibile esprimere l'intersezione tramite negazioni ed unioni di eventi.

∘ Anche se vi è una corrispondenza formale, detta *isomorfismo*, fra «teoria degli insiemi» ed «algebra degli eventi» (il che agevola comprensione e rappresentazione degli eventi e delle relazioni tra di essi), va sottolineato che *gli eventi non coincidono con gli insiemi*. Quindi, non si deve utilizzare la terminologia della teoria degli insiemi per il Calcolo delle probabilità.

Le operazioni tra eventi godono di numerose proprietà, tutte mutuate dall'isomorfismo con l'algebra degli insiemi, ed agevolmente memorizzabili perché ovvie ed immediate. L'elenco seguente è limitato alle proprietà più utili; si noti che le proprietà associativa, distributiva e di idempotenza sono generalizzabili ad un numero qualsiasi di eventi.

Proprietà	Unione	Intersezione
Idempotenza	$A \cup A = A$	$A \cap A = A$
Elemento neutro	$A \cup \emptyset = A$	$A \cap \Omega = A$
Commutativa	$A \cup B = B \cup A$	$A \cap B = B \cap A$
Associativa	$(A \cup B) \cup C = A \cup (B \cup C)$	$(A \cap B) \cap C = A \cap (B \cap C)$
Distributiva	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$	$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

Una rappresentazione grafica degli eventi, particolarmente efficace, è quella dei *diagrammi di Venn*, generalmente proposta per gli insiemi di punti del piano.

Lo spazio campione Ω costituisce la totalità degli eventi di una prova e viene, quindi, disegnato mediante una figura chiusa (generalmente un rettangolo) all'interno della quale si collocano le diverse figure chiuse (di dimensioni e forme arbitrarie) le quali rappresentano gli eventi.

Su tali diagrammi, l'unione tra due eventi A e B sarà evidentemente rappresentata dalla «parte complessiva sia di A che di B»; l'intersezione come la «parte in comune, se esiste, tra la rappresentazione di A e quella di B»; la negazione di A come la «parte di Ω non appartenente alla rappresentazione di A», e così via.

∘ Naturalmente, *i diagrammi di Venn non dimostrano alcuna relazione* perché sono solo uno strumento di agevole rappresentazione grafica delle relazioni e delle operazioni sull'algebra degli eventi. Infatti, ciò che è importante nei diagrammi di Venn non è l'esatta specificazione delle figure utilizzate per rappresentare gli eventi ma la reciproca collocazione, perché agevola il riferimento alle proprietà formali degli eventi.

La Figura 8.1 conferma come la terminologia utilizzata per l'algebra degli eventi sia in perfetta sintonia con la rappresentazione tramite i diagrammi di Venn. Per esempio,

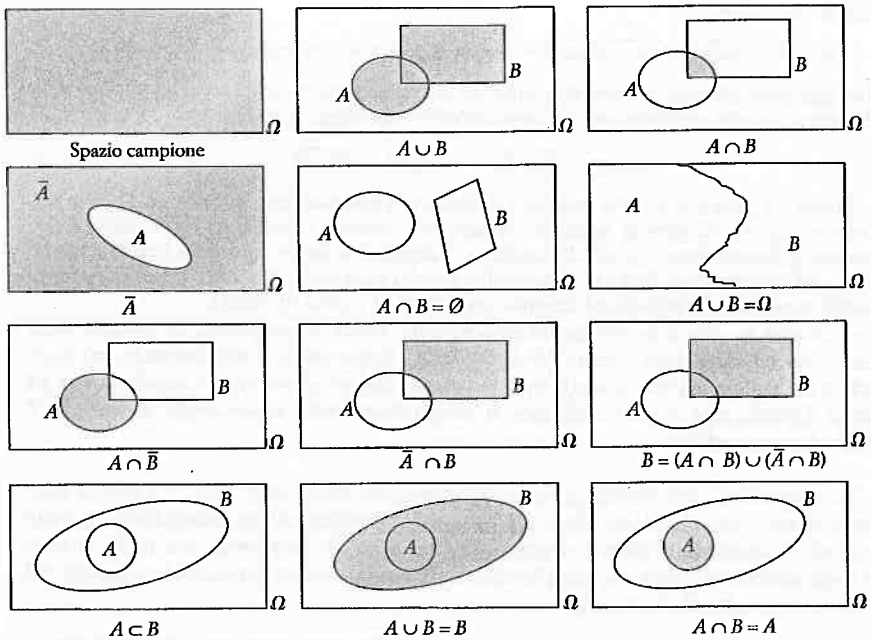

FIGURA 8.1. Diagrammi di Venn e relazioni fondamentali tra eventi.

l'evento $A \cap \overline{B}$, che si legge: «A e non B» è rappresentato, alla lettera, dalla «parte di A che non appartiene a B», e così via.

Grazie ai diagrammi di Venn, le definizioni seguenti, che possono essere poste in relazione ad eventi generici, diventano di significato immediato.

• Due eventi si dicono ***incompatibili*** se non possono verificarsi contemporaneamente. Quindi, A e B sono eventi incompatibili se: $A \cap B = \emptyset$.

○ Sui diagrammi di Venn, la rappresentazione di due eventi incompatibili sarà quella di due porzioni di piano che non hanno alcuna parte in comune.

○ Data una collezione di eventi $E_i \in \mathcal{A}, i = 1, 2, \ldots$ diremo che essi sono *incompatibili a due a due* se qualunque coppia di eventi distinti non può verificarsi contemporaneamente, cioè se: $E_i \cap E_j = \emptyset$, $\forall i \neq j$. Evidentemente, se fosse $i = j$ sarebbe sempre: $E_i \cap E_i = E_i$ per la proprietà di idempotenza della intersezione.

○ Analogamente, parleremo di *incompatibilità a tre a tre,..., ad m ad m, ...* se la intersezione di qualsiasi gruppo di tre, ..., m,... eventi distinti dà luogo all'evento impossibile.

○ È agevole dedurre che la incompatibilità a due a due implica tutte le incompatibilità di ordine superiore, ma non viceversa. Quando in seguito diremo *eventi incompatibili*, senza alcuna specificazione, intenderemo sempre «eventi incompatibili a due a due».

• Due eventi A e B si dicono ***necessari*** se la loro unione è l'evento certo,

cioè se si verifica certamente almeno uno di essi. Quindi, due eventi A e B sono necessari se: $A \cup B = \Omega$.

○ Sui diagrammi di Venn, la rappresentazione di due eventi necessari consiste in due porzioni del piano che «insieme ricoprono completamente» lo spazio Ω.

Una collezione di eventi $E_i, i = 1, 2, \ldots$ *si dice necessaria* (oppure, che gli eventi E_i sono necessari) *se almeno uno di essi certamente si verifica*, cioè se: $\bigcup E_i = \Omega$.

• Se una collezione di eventi è composta da *eventi necessari e incompatibili*, allora essa è una *partizione* di Ω.

Formalmente, la collezione di eventi $E_i, i = 1, 2, \ldots$ appartenenti ad una prefissata prova costituisce una ***partizione*** di Ω se e solo se:

i) $\bigcup E_i = \Omega$ (*necessarietà*);

ii) $E_i \cap E_j = \emptyset$, $\forall i \neq j$ (*incompatibilità a due a due*).

○ Sui diagrammi di Venn, una partizione di eventi si rappresenta come una collezione di figure che ricoprono Ω senza sovrapporsi; quindi, costituiscono una «suddivisione in parti» di Ω. In senso figurato, si potrebbe anche dire che una partizione di Ω è un «puzzle» di eventi.

Esempio 8.6

Nella prova «estrazione di una pallina da un'urna che contiene 20 palline numerate da 1 a 20» si ha: $\Omega = \{E_1, E_2, \ldots, E_{20}\}$. Definiamo, quindi, i seguenti eventi:

$$A = \{E_1, E_2, E_3, E_5, E_6, E_7\}; \quad B = \{E_7, E_8, E_{11}, E_{12}, E_{15}, E_{16}\};$$
$$C = \{E_{13}, E_{14}, E_{15}, E_{17}, E_{18}, E_{19}\}; \quad D = \{E_4, E_9, E_{10}, E_{20}\}.$$

Tali eventi sono necessari perché: $A \cup B \cup C \cup D = \Omega$, ma essi non costituiscono una partizione di Ω, perché non sono incompatibili a due a due. In effetti, si ha:

$$A \cap C = A \cap D = B \cap D = C \cap D = \emptyset \; ; \; \textit{ma} \; A \cap B = \{E_7\}; \; B \cap C = \{E_{15}\}.$$

Si osservi che, in questa prova, gli eventi A, B e C sono incompatibili a 3 a 3 perché: $A \cap B \cap C = \emptyset$; tuttavia, non sono incompatibili a 2 a 2. □

• Una importante relazione tra eventi è la ***inclusione*** che si definisce nel modo seguente. Un evento A è incluso nell'evento B se tutte le volte che si verifica A certamente si verifica B; scriveremo: $A \subset B$ (e si legge «A incluso in B»). Se A è incluso in B, si può anche affermare che: «B include A», il che si indica con: $B \supset A$.

○ La terminologia deriva dal fatto che l'evento $A \subset B$ si rappresenta nei diagrammi di Venn proprio mediante la inclusione geometrica della rappresentazione di A nella rappresentazione di B.

o La inclusione tra due eventi A e B non esclude la perfetta coincidenza tra A e B; inoltre, se A è incluso in B e B è incluso in A, allora i due eventi coincidono; e viceversa. Formalmente:

$$A \subset B\,,\, B \subset A \Longleftrightarrow A \equiv B.$$

o La relazione di inclusione è transitiva:

$$A \subset B\,,\, B \subset D \Longrightarrow A \subset D.$$

Inoltre, si può dimostrare che se A è incluso in B, allora $\overline{B}$ è incluso in $\overline{A}$, e viceversa:

$$A \subset B \Longleftrightarrow \overline{B} \subset \overline{A}.$$

o Sul piano formale, condizioni necessarie e sufficienti perché l'evento A sia incluso nell'evento B sono il verificarsi delle due relazioni:

$$A \cup B = B;\ A \cap B = A \Longleftrightarrow A \subset B.$$

o Tramite esse è possibile dimostrare che, per qualsiasi evento E, si ha:

$$\emptyset \subset E \subset \Omega.$$

Infatti, se si pone $A = \emptyset\ ; B = E$, risulta: $\emptyset \cup E = E$; $\emptyset \cap E = \emptyset$ e, quindi, $\emptyset \subset E$. Allo stesso modo, se si pone: $A = E; B = \Omega$, risulta: $E \cup \Omega = \Omega; E \cap \Omega = E$ e, quindi, $E \subset \Omega$.

8.9 Postulati del Calcolo delle probabilità

Presentiamo, ora, i postulati proposti da Kolmogorov secondo la formalizzazione corrente, ricordando che ad essi andrebbe premesso il postulato secondo cui «gli eventi formano un'algebra di Boole completa».

Siano $E_i, i = 1, 2, \ldots$ eventi di Ω. *La **probabilità** di un evento E è una funzione di insieme a valori reali* (che indicheremo con $Pr(E)$ e leggeremo «probabilità di E») che soddisfa i seguenti postulati:

Postulato 1. $Pr(E_i) \geq 0, \forall\, E_i \subset \Omega.$

Postulato 2. $Pr(\Omega) = 1.$

Postulato 3. $Pr(E_i \cup E_j) = Pr(E_i) + Pr(E_j)$, se $E_i \cap E_j = \emptyset\ , \forall i \neq j.$

o In modo discorsivo, i postulati affermano che:

Postulato 1. La probabilità è una funzione di insieme che assegna ad ogni elemento di Ω (ossia a qualsiasi evento E_i) un numero reale non negativo (cioè maggiore o uguale a 0).

Postulato 2. L'evento certo Ω ha probabilità 1.

Postulato 3. La probabilità della unione di due *eventi incompatibili*, cioè tali che $E_i \cap E_j = \emptyset$, è la somma delle probabilità dei singoli eventi.

o Il terzo postulato si generalizza agevolmente alla somma di tre, quattro, ..., m eventi tra loro incompatibili, affermando che: *la probabilità dell'unione di un numero qualsiasi di eventi tra loro incompatibili è la somma delle probabilità dei singoli eventi.*

In effetti, è possibile sostituire il terzo postulato con l'affermazione che la probabilità dell'unione anche infinita numerabile di eventi incompatibili è la somma infinita (=serie) delle singole probabilità, ed in tal caso si parla di «sommabilità completa». In questo volume, quando occorre, si supporrà che il terzo postulato sia valido per un numero qualsiasi di eventi incompatibili, anche infinito numerabile.

Quindi, per noi, sarà sempre valido il terzo postulato nella seguente estensione della *sommabilità completa*:

$$\textbf{\textit{Postulato 3.bis}}\ Pr\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} Pr(E_i)\,,\ \textit{se}\ E_i \cap E_j = \emptyset\ , \forall i \neq j.$$

o Ovviamente, i postulati non possono essere dimostrati ma su di essi vi è un accordo generalizzato (sia da parte dei frequentisti che da parte dei soggettivisti, che però limitano il terzo postulato ad un numero finito di eventi incompatibili). Infatti, le tre affermazioni precedenti si accordano sia con le proprietà delle frequenze relative che con quelle della somma che una persona coerente è disposta a scommettere sul verificarsi degli eventi.

8.10 Principali teoremi del Calcolo delle probabilità

Grazie ai postulati si possono derivare numerosi teoremi, a loro volta utilizzabili per dimostrare altri teoremi. Quelli che presenteremo sono i più semplici ed utili.

▶ *Teorema 8.1* $Pr(B \cap \overline{A}) = Pr(B) - Pr(A \cap B).$

▶ *Teorema 8.2* $Pr(\overline{A}) = 1 - Pr(A).$

▶ *Teorema 8.3* $Pr(\emptyset) = 0.$

▶ *Teorema 8.4* $A \subset B \Longrightarrow Pr(A) \leq Pr(B).$

▶ *Teorema 8.5* $\forall\, E \subset \Omega,\ 0 \leq Pr(E) \leq 1.$

▶ *Teorema 8.6* $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B).$

▶ *Teorema 8.7* $Pr(A \cup B \cup C) = Pr(A) + Pr(B) + Pr(C) - Pr(A \cap B) - Pr(A \cap C) - Pr(B \cap C) + Pr(A \cap B \cap C).$

Prima di accennare ad una loro dimostrazione, esplicitiamo alcuni commenti ai precedenti teoremi per una migliore comprensione ed applicazione.

o Il teorema 8.1 afferma che la probabilità dell'evento «si verifica B ma non A» è uguale alla probabilità di B meno la probabilità di A e B.

o Il teorema 8.2 afferma che la probabilità dell'evento $\overline{A}$ è il complemento ad 1 della probabilità dell'evento A.

o Il teorema 8.3 afferma che la probabilità dell'evento impossibile è 0. Si osservi a tal proposito che questo è un *teorema*, per cui esso si dimostra rigorosamente a partire dai postulati, mentre l'affermazione per cui la probabilità dell'evento certo è 1 costituisce un *postulato* per cui non si può dimostrare. Inoltre, mentre $\emptyset$ è un evento, sottoposto alle regole dell'algebra di Boole, la sua probabilità, cioè 0 è un *numero*, sottoposto alle regole dell'aritmetica. Infatti, bisogna tenere ben distinti l'evento $\emptyset$ e il numero 0 (che è la probabilità dell'evento impossibile $\emptyset$).

o Il teorema 8.4 afferma che se un evento è incluso in un altro, la probabilità del primo è sempre non superiore alla probabilità del secondo.

o Il teorema 8.5 afferma che la probabilità di qualsiasi evento è sempre un numero compreso tra 0 e 1. È importante notare che tale affermazione, per quanto intuitiva e prevedibile, è un *teorema*, cioè una affermazione che si dimostra rigorosamente a partire dai postulati.

o Il teorema 8.6 esplicita la probabilità della unione di due eventi qualsiasi (non necessariamente incompatibili, quindi) come la somma delle singole probabilità meno la probabilità della loro intersezione. Ovviamente, se gli eventi sono incompatibili, poiché l'intersezione è impossibile e l'evento impossibile ha probabilità 0 (per il teorema 8.3), ci si ritrova il terzo postulato. Ovviamente, *questo teorema non dimostra il terzo postulato* perché tale postulato è un presupposto essenziale per la dimostrazione di questo stesso teorema.

o Il teorema 8.7 estende il teorema 8.6 al caso della probabilità della unione di 3 eventi qualsiasi. È possibile generalizzare tale teorema, perché si dimostra che la probabilità della unione di un numero qualsiasi di eventi è uguale alla somma delle singole probabilità meno la somma di tutte le probabilità delle intersezioni a 2 a 2 più la somma di tutte le probabilità delle intersezioni a 3 a 3 meno la somma di tutte le probabilità delle intersezioni a 4 a 4,..., sino a più (o meno) la probabilità della intersezione di tutti gli eventi.

Formalmente, esplicitiamo la probabilità della unione di un numero m qualsiasi di eventi:

$$Pr\left(\bigcup_{i=1}^{m} E_i\right) = \sum_{i=1}^{m} Pr(E_i) - \sum_{i=1}^{m}\sum_{j=1}^{m} Pr(E_i \cap E_j) + \\ + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} Pr(E_i \cap E_j \cap E_k) - \ldots (\pm 1)^{m-1} Pr\left(\bigcap_{i=1}^{m} E_i\right).$$

Non dimostreremo i precedenti teoremi, eccetto il primo dal quale si deducono, poi, i successivi due. Lo stile della dimostrazione è comunque simile per tutti perché si sfruttano le relazioni tra eventi ed i postulati, nonché i teoremi di volta in volta già dimostrati.

o Per dimostrare il teorema 8.1, osserviamo che un qualsiasi evento B si può esprimere come l'unione dei 2 eventi incompatibili $(B \cap A)$ e $(B \cap \overline{A})$. Ciò è evidente dai diagrammi di Venn ma è anche confermato sul piano logico: un evento B si verifica assieme ad A, e si verifica allora $(B \cap A)$, oppure si verifica senza A, e si verifica $(B \cap \overline{A})$. Quindi, sarà sempre valida la relazione:

$$B = (B \cap A) \cup (B \cap \overline{A}),$$

nella quale, essendo: $(B \cap A) \cap (B \cap \overline{A}) = \emptyset$ i due eventi in parentesi sono fra loro incompatibili.

È possibile allora applicare il terzo postulato ed ottenere:

$$Pr(B) = Pr\left((B \cap A) \cup (B \cap \overline{A})\right) = Pr(B \cap A) + Pr\left(B \cap \overline{A}\right).$$

Esplicitando al primo membro $Pr\left(B \cap \overline{A}\right)$ si ottiene il teorema 8.1, che è così dimostrato.

o Ora, se nel teorema 8.1 appena dimostrato, si pone $B = \Omega$, grazie al secondo postulato si dimostra immediatamente il teorema 8.2.

o Ancora, se nel teorema 8.2 si pone $A = \Omega$ si dimostra subito il teorema 8.3.

8.11 Probabilità condizionata e indipendenza

Talvolta occorre calcolare la probabilità di un evento dopo aver constatato che si è verificato un altro evento, logicamente o temporalmente collegato al primo. Questa è una situazione assai comune negli esperimenti sequenziali perché, in tali casi, la decisione finale dipende dai risultati e dalle decisioni intermedie. Gli esempi seguenti mostrano delle prove nelle quali i risultati di alcune sottoprove possono modificare oppure no la probabilità dei risultati successivi.

Esempio 8.7

o Nella prova «lancio di più monete» è ragionevole affermare che qualsiasi evento si verifichi in un lancio non modifica la probabilità che si verifichi un dato evento in un altro lancio, sia che tali lanci siano effettuati consecutivamente che contemporaneamente.

o Nella prova «estrazione di due palline *con ripetizione*» da un'urna che contiene 1 pallina Bianca e 9 Nere, supponiamo che la vincita avviene se almeno una delle due palline è quella Bianca. Evidentemente, qualsiasi sia il risultato della prima pallina estratta esso non modifica la situazione della seconda estrazione, perché il rimettere la pallina estratta nell'urna non modifica la probabilità degli eventi connessi alla seconda estrazione: la seconda sottoprova è in effetti una replica della prima.

Per contro, se l'estrazione viene effettuata *senza ripetizione*, cioè procedendo all'estrazione della seconda pallina senza rimettere nell'urna la prima pallina estratta, allora se la prima pallina estratta è Bianca la probabilità che anche la seconda sia Bianca è 0 perché è impossibile che essa venga ri-estratta. D'altra parte, se la prima pallina estratta è Nera, la probabilità che la seconda pallina sia Bianca è cambiata poiché ora nell'urna non vi sono più 10 palline ma solo 9.

o Nella prova «lancio di un dado ben equilibrato» è ragionevole attribuire una eguale probabilità ai 6 eventi elementari $\{E_1, E_2, \ldots, E_6\} = \Omega$ connessi alle 6 facce (secondo l'impostazione classica discussa nel paragrafo 8.4). Ma se il banco informa il giocatore che si è verificato «un numero pari», nessuno sarà disposto a scommettere, per esempio, sull'evento E_3. Infatti, lo spazio campionario Ω si è ora ridotto allo spazio campionario $\Omega' = \{E_2, E_4, E_6\}$ consistente in 3 soli eventi elementari (quelli connessi alle facce con un numero pari di punti). Il giocatore, ricevuta l'informazione che si è verificato un numero pari, modifica le sue valutazioni probabilistiche sugli eventi di Ω, attribuendo ad alcuni eventi probabilità 0 (perché, ora, sono impossibili) e modifica conseguentemente anche la probabilità di quelli rimanenti (perché, ora, non si tratta più di scommettere su 6 eventi possibili ma solo su 3). □

La conseguenza più evidente del verificarsi di un evento A sulla probabilità di un evento B è il fatto che, ora, lo spazio campione Ω è diventato A; l'evento A assume il ruolo di evento certo (infatti, si è verificato) e qualunque evento che ha probabilità maggiore di 0 deve essere incluso in A. Quindi, se A si è verificato, gli eventi elementari che compongono B devono necessariamente essere inclusi anche in A, cioè sono gli eventi elementari comuni a B e ad A, cioè $A \cap B$.

o Appare quindi ragionevole, qualsiasi sia l'interpretazione della probabilità cui si aderisce, considerare la «probabilità dell'evento B dato che l'evento A si è verificato» come il rapporto tra la probabilità dell'evento $A \cap B$ e la probabilità dell'evento A.

Pertanto, si definisce la ***probabilità condizionata*** di B dato A, che si scrive $Pr(B|A)$, nel modo seguente:

$$\text{Se } Pr(A) > 0, \text{ allora: } \quad Pr(B|A) = \frac{Pr(A \cap B)}{Pr(A)} .$$

o La probabilità condizionata appena definita soddisfa tutti i postulati del Calcolo delle probabilità.

o La probabilità condizionata ha senso solo se il denominatore della precedente definizione è diverso da 0; ecco perché è necessario premettere il vincolo: $Pr(A) > 0$.

o È agevole constatare che la probabilità tradizionale $Pr(E)$ di qualsiasi evento $E \subset \Omega$ può essere sempre interpretata come una probabilità condizionata, perché: $Pr(E|\Omega) \equiv Pr(E)$.

Introducendo la probabilità condizionata, è ora possibile esprimere la probabilità della intersezione di due (o più) eventi tramite le probabilità dei singoli eventi. Infatti, dalla definizione, si trae subito:

$$Pr(A \cap B) = Pr(A)\ Pr(B|A) .$$

o Tale relazione esprime la probabilità del contemporaneo presentarsi di due eventi come il prodotto della probabilità del primo evento per la probabilità del secondo evento dato che il primo evento si è già presentato. Ovviamente, poiché $A \cap B = B \cap A$, dalla definizione di probabilità condizionata segue anche: $Pr(A \cap B) = Pr(B)\ Pr(A|B)$.

Tali risultati si possono generalizzare ad una successione di eventi $E_1, E_2, E_3 \ldots$, mediante la relazione seguente:

$$Pr(E_1 \cap E_2 \cap E_3 \cap \ldots) = Pr(E_1)\ Pr(E_2|E_1)\ Pr(E_3|E_1 \cap E_2) \ldots$$

Nell'esempio 8.7 sono state descritte alcune prove nelle quali la probabilità di un certo evento viene modificata in conseguenza del verificarsi di altri eventi; la formalizzazione di questi concetti conduce alla definizione di eventi e sottoprove indipendenti.

Sul piano concettuale, due eventi A e B, inclusi in Ω, si dicono *indipendenti* se il verificarsi dell'evento A non modifica la probabilità dell'evento B cioè se, essendo $Pr(A) > 0$, si ha: $Pr(B|A) = Pr(B)$. Ora, è immediato derivare – sostituendo

$Pr(B|A)$ nella definizione di probabilità condizionata – che se $Pr(B|A) = Pr(B)$ allora: $Pr(B) = \dfrac{Pr(A \cap B)}{Pr(A)}$.

Pertanto, sul piano formale si giunge alla seguente definizione:

• *Due* ***eventi*** *A e B si dicono* ***indipendenti*** *se e solo se:*

$$Pr(A \cap B) = Pr(A)\ Pr(B) .$$

Dalle relazioni precedenti, si può anche dimostrare il seguente teorema:

▶ *Teorema 8.8* In una certa prova, se A e B sono eventi indipendenti, allora lo sono anche $\overline{A}$ e B, A e $\overline{B}$, $\overline{A}$ e $\overline{B}$.

o Va sottolineato che ciò che è veramente *condizionato* è la probabilità, e non l'evento coinvolto. In effetti, $A, B, \ldots$ restano tali ma il verificarsi di uno di essi modifica la misura della probabilità che noi attribuiamo al loro verificarsi. Per questo, l'indipendenza è una relazione tra probabilità e *non* tra eventi: il verificarsi di A, infatti, non modifica *la probabilità di B* se sono indipendenti (non ha alcun senso la frase: «non modifica B», perché ciò che si modifica è un numero e non un evento).

o Il concetto di indipendenza si può estendere. Data una collezione di eventi $E_1, E_2, \ldots$, tutti inclusi in Ω, si dirà che essi sono *mutuamente indipendenti* se, per ogni gruppo di $m > 1$ eventi, si ha:

$$Pr(E_1 \cap E_2 \cap \cdots \cap E_m) = Pr(E_1)\ Pr(E_2) \ldots Pr(E_m) , \ \forall m = 2, 3, \ldots$$

Quindi, perché si possa parlare di mutua indipendenza di una collezione di eventi occorre che tutti gli eventi siano indipendenti a 2 a 2, a 3 a 3, ..., a m a m, per qualsiasi m; per contro, l'indipendenza a 2 a 2 non implica necessariamente quella a 3 a 3.

o Essendo molto diffusa una certa confusione tra incompatibilità e indipendenza degli eventi, è opportuno ripetere e confrontare i due concetti in modo schematico, come nella tabella seguente.

	Incompatibilità	***Indipendenza***
Definizione	$A \cap B = \emptyset$	$Pr(A \cap B) = Pr(A)\ Pr(B)$
Relazione	tra eventi	tra probabilità
Rappresentazione	sui diagrammi di Venn	non si rappresenta
Conseguenza	su $Pr(A \cup B)$	su $Pr(A \cap B)$

D'altra parte, i termini probabilistici richiamano quelli del linguaggio corrente. Infatti, si parla di persone che hanno tra loro caratteri «incompatibili» (quando le persone hanno difficoltà a stare assieme) e di persone che sono «indipendenti» (quando hanno autonomia di giudizio, comportamento, reddito, etc.).

In breve, *la incompatibilità è una relazione tra eventi*, che ha come conseguenza il fatto che la probabilità della unione di più eventi incompatibili è la somma delle singole probabilità. Invece, *la indipendenza è una relazione tra le probabilità* degli eventi ed ha come conseguenza il fatto che la probabilità della intersezione di più eventi è il prodotto delle singole probabilità. Nello schema seguente, che esplicitiamo per 2 soli eventi A e B, riassumiamo queste relazioni il cui impatto è importante nella successiva misura della probabilità di eventi composti.

$$\begin{array}{rcl} A \text{ e } B \textit{ eventi INCOMPATIBILI} & \Longrightarrow & P_r(A \cup B) = P_r(A) + P_r(B) \\ A \text{ e } B \textit{ eventi NON INCOMPATIBILI} & \Longrightarrow & P_r(A \cup B) = P_r(A) + P_r(B) - P_r(A \cap B) \\ A \text{ e } B \textit{ eventi INDIPENDENTI} & \Longrightarrow & P_r(A \cap B) = P_r(A)\, P_r(B) \\ A \text{ e } B \textit{ eventi NON INDIPENDENTI} & \Longrightarrow & P_r(A \cap B) = P_r(A)\, P_r(B|A) \end{array}$$

Un risultato importante concernente le probabilità condizionate è il seguente ***teorema delle probabilità totali***, che ci limitiamo ad esplicitare senza dimostrazione.

▶ *Teorema 8.9* Sia $E_1, E_2, \ldots$ una partizione di eventi di Ω. Allora, per ogni evento $A \subset \Omega$, si ha:

$$P_r(A) = \sum_{i=1}^{\infty} P_r(E_i)\, P_r(A|E_i)\ .$$

Il teorema precedente consente una interpretazione interessante circa il ruolo delle probabilità condizionate. Infatti, poiché gli eventi $E_1, E_2, \ldots$ sono una partizione di Ω, allora la somma delle loro probabilità è 1. Quindi, il teorema delle probabilità totali afferma che la probabilità di qualsiasi altro evento $A \subset \Omega$ è pari alla media aritmetica ponderata delle probabilità condizionate $P_r(A|E_i)$ con pesi $P_r(E_i)$.

8.12 La misura della probabilità

Nelle pagine precedenti sono state discusse numerose affermazioni concernenti gli eventi e le loro probabilità *senza mai pervenire alla misura di tali probabilità* (eccezion fatta per gli eventi $\emptyset$ e Ω). In effetti, a differenza delle impostazioni classica, frequentista e soggettivista (ove si introduce la probabilità come numero associato al presentarsi di un evento proponendone anche la misura), nella assiomatizzazione il problema della misura delle probabilità non è prioritario. Successivamente, sulla base delle caratteristiche della prova (e, quindi, della struttura di Ω), si assegnerà una misura per la probabilità che non contraddica i Postulati.

o Per motivi di semplicità, anche se i risultati si estendono al caso continuo, limiteremo la discussione alle prove che generano partizioni discrete (cioè i cui eventi sono elencabili in un numero finito o numerabile), fornendo solo qualche accenno alla misura della probabilità nei casi di prove che generano partizioni continue (cioè quelle i cui eventi sono in corrispondenza di tutti i punti di un intervallo reale).

Supponiamo, quindi, che la prova $\mathcal{P}$ generi un numero finito n di eventi elementari $E_1, E_2, \ldots, E_n$, i quali costituiscono una partizione dello spazio campione Ω. Inoltre, supponiamo che, per ragioni logiche, geometriche, fisiche, psicologiche o altre (derivanti dalla particolare struttura della prova o dalla conoscenza che si possiede su di essa), tali eventi siano equiprobabili, cioè la probabilità di E_i sia un numero costante al variare di $i = 1, 2, \ldots, n$ (anche se non conosciamo questo specifico valore).

Allora, la classe di eventi *necessari, incompatibili ed equiprobabili* $\{E_i, i = 1, 2, \ldots, n\}$ si definisce una ***partizione equiprobabile*** di eventi di Ω.

Formalmente, una *partizione equiprobabile* di eventi di Ω è una collezione di eventi $E_1, E_2, \ldots, E_n$ tali che:

i) $E_1 \cup E_2 \cup \cdots \cup E_n = \Omega$; $\Longleftarrow$ *necessarietà*;

ii) $E_i \cap E_j = \emptyset$, $\forall i \neq j$; $\Longleftarrow$ *incompatibilità*;

iii) $P_r(E_1) = P_r(E_2) = \cdots = P_r(E_n)$; $\Longleftarrow$ *equiprobabilità*.

Se in una prova gli eventi elementari costituiscono una partizione equiprobabile è possibile assegnare in modo univoco la misura della probabilità a qualsiasi evento generato dalla prova derivandola direttamente dai Postulati.

La giustificazione formale è immediata e deriva dai seguenti sviluppi:

$$\begin{array}{lll} 1 & = P_r(\Omega) = & \text{(per il Postulato 2)} \\ & = P_r(E_1 \cup E_2 \cup \cdots \cup E_n) = & \text{(perché gli eventi sono necessari)} \\ & = P_r(E_1) + P_r(E_2) + \cdots + P_r(E_n) = & \text{(perché gli eventi sono incompatibili)} \\ & = P_r(E_i) + P_r(E_i) + \cdots + P_r(E_i) = & \text{(perché gli eventi sono equiprobabili)} \\ & = n\, P_r(E_i)\,. & \end{array}$$

Da tale catena di uguaglianze si deduce: $1 = n\, P_r(E_i)$ e, quindi:

$$\boxed{P_r(E_i) = \frac{1}{n}\ ,\ \forall i = 1, 2, \ldots, n.}$$

In particolare, se un evento $A \subset \Omega$ è composto di $m \leq n$ eventi elementari, cioè se $A = E_1 \cup E_2 \cup \ldots E_m$, allora:

$$\begin{array}{rcl} P_r(A) & = & P_r(E_1 \cup E_2 \cup \ldots E_m) = P_r(E_1) + P_r(E_2) + \cdots + P_r(E_m) \\ & = & \underbrace{\frac{1}{n} + \frac{1}{n} + \cdots + \frac{1}{n}}_{m \text{ volte}} = \frac{m}{n}. \end{array}$$

Quest'ultimo risultato dimostra che «*in una partizione equiprobabile di eventi appartenenti allo spazio Ω, la probabilità di qualsiasi evento $A \subset \Omega$ è il rapporto tra il numero degli eventi elementari di cui A si compone* (casi favorevoli ad A) *ed il numero totale degli eventi elementari di Ω* (casi ugualmente possibili)».

Quindi, *solo nel caso di una partizione equiprobabile*, la misura della probabilità si risolve nella *regola pratica* che calcola il rapporto tra casi favorevoli all'evento e casi possibili (come sosteneva la definizione classica: paragrafo 8.4). Tuttavia, si osservi che essa *non* costituisce una definizione di probabilità ma solo una sua misura in una prova molto particolare, quella cioè che genera *una partizione equiprobabile di eventi elementari*. E tale misura è una assegnazione della probabilità rigorosamente derivata dai Postulati.

Grazie a tale regola, è possibile calcolare la probabilità di eventi generati da prove nelle quali è ragionevole assumere la equiprobabilità degli eventi elementari.

Come l'esempio mostrerà, per numerose e differenti situazioni sperimentali, a partire da tale regola, sarà anche possibile pervenire a generalizzazioni successive.

Esempio 8.8

• Nella prova «lancio di due monete ben equilibrate», supponiamo di voler calcolare la probabilità dell'evento E =«almeno una delle due monete è Testa». Lo spazio Ω dei possibili risultati, con ovvia simbologia, risulta essere $\Omega = \{E_1, E_2, E_3, E_4\}$, dove:

$$E_1 = (T,T); \quad E_2 = (T,C); \quad E_3 = (C,T); \quad E_4 = (C,C).$$

Tali eventi sono *necessari* (certamente si verifica uno di essi), *incompatibili* (nessuno dei 4 eventi può verificarsi congiuntamente ad un altro) ed *equiprobabili* (perché essendo le monete ben bilanciate non vi è alcuna ragione per attribuire ad uno dei quattro eventi una probabilità maggiore o minore di quella degli altri). Quindi, Ω è una partizione equiprobabile di $n = 4$ eventi elementari per cui, secondo la regola appena dimostrata, ciascuno di essi ha probabilità:

$$Pr(E_1) = Pr(E_2) = Pr(E_3) = Pr(E_4) = \frac{1}{4}.$$

Ora, l'evento E =«almeno una delle due monete è Testa» si verifica se, nella prova, si verifica uno dei seguenti eventi elementari: E_1, E_2 oppure E_3, per cui $E = E_1 \cup E_2 \cup E_3$; pertanto, la probabilità di E sarà ottenuta da:

$$Pr(E) = Pr(E_1 \cup E_2 \cup E_3) = Pr(E_1) + Pr(E_2) + Pr(E_3) = \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{3}{4}.$$

Tale risultato poteva anche dedursi immediatamente dalla regola pratica prima enunciata. Infatti, nell'ambito della partizione equiprobabile descritta, l'evento E è composto di 3 eventi elementari su un totale di 4 eventi complessivi per cui il «rapporto tra casi favorevoli ad E e casi possibili» fornisce 3/4 come probabilità di E.

• Nella prova «lancio di due dadi simmetrici», i 36 eventi elementari (E_i, E_j), per $(i,j) = 1,2,\ldots 6$, che descrivono lo spazio Ω, formano una partizione equiprobabile di eventi a ciascuno dei quali si assegna probabilità pari a 1/36.

Allora, l'evento E =«la somma dei punti sulle due facce è 5» si verifica se e solo se si verificano gli eventi $(E_1, E_4), (E_2, E_3), (E_3, E_2), (E_4, E_1)$. Pertanto, la probabilità richiesta è:

$$Pr(E) = \frac{4}{36} = \frac{1}{9}.$$

• Nella prova «gioco di una schedina al Totocalcio», la probabilità di vincere un «tredici» giocando una sola colonna viene ottenuta facendo il rapporto tra l'unica colonna giocata e le $3^{13} = 1.594.323$ possibili colonne che possono verificarsi. Tale probabilità è $1/3^{13} \simeq 0.62723 * 10^{-6}$. In effetti, questo calcolo è corretto solo se si assume che tutte le colonne siano equiprobabili, il che è certamente falso nel gioco del Totocalcio (mentre è corretto nel gioco dell'Enalotto). Infatti, nessuno è disposto ad attribuire la stessa probabilità ad una colonna con tutti «segni 2» rispetto ad una colonna con «7 segni 1, 4 segni X, 2 segni 2», e così via.

• L'estrazione del Lotto è effettuata in 10 città italiane (dette «ruote») scegliendo 5 numeri *senza ripetizione* da un'urna che contiene gli interi da 1 a 90. La probabilità di vincere puntando un solo numero su una ruota prefissata è: $5/90 = 1/18 = 0.05556$.

Ovviamente, il giocatore può puntare lo stesso numero su tutte le ruote. Indicando con E_i l'evento «il numero esce sulla ruota i-esima», la probabilità che il numero sia estratto in almeno una ruota è: $Pr(E_1 \cup E_2 \cup \cdots \cup E_{10})$. Essa può essere calcolata, sfruttando le leggi di de Morgan e l'indipendenza tra le 10 sottoprove, nel modo seguente:

$$\begin{aligned} Pr(E_1 \cup E_2 \cup \cdots \cup E_{10}) &= Pr\left(\overline{\overline{E}_1 \cap \overline{E}_2 \cap \cdots \cap \overline{E}_{10}}\right) \\ &= 1 - Pr\left(\overline{E}_1 \cap \overline{E}_2 \cap \cdots \cap \overline{E}_{10}\right) \\ &= 1 - Pr\left(\overline{E}_1\right) Pr\left(\overline{E}_2\right) \ldots Pr\left(\overline{E}_{10}\right) \\ &= 1 - \left(1 - \frac{1}{18}\right)\left(1 - \frac{1}{18}\right) \cdots \left(1 - \frac{1}{18}\right) \\ &= 1 - \left(1 - \frac{1}{18}\right)^{10} = 0.43537 \end{aligned}$$

• Nella prova «estrazione di una pallina» da un'urna che contiene 3 palline Verdi e 7 Palline Rosse, evidentemente, gli eventi: V =«la pallina estratta è Verde» e R =«la pallina estratta è Rossa», pur costituendo una partizione di Ω (infatti, sono necessari e incompatibili), *non sono equiprobabili* per cui non si può applicare la regola pratica prima enunciata per il calcolo delle probabilità di R e V.

Tuttavia, se si considerano le *singole* $3 + 7 = 10$ palline, le quali a prescindere dal colore hanno tutte la stessa forma, peso, rugosità, etc. allora l'estrazione di ciascuna delle 10 palline genera gli eventi $\{E_1, E_2, \ldots, E_{10}\}$ i quali costituiscono una partizione equiprobabile di Ω: infatti, non si vede perché una delle 10 palline debba avere una probabilità diversa dalle altre essendo le 10 palline tra loro indistinguibili.

Allora, per esempio, si può supporre che le palline Verdi siano le prime 3 e le palline Rosse siano le ultime 7, in modo che la prova consista degli eventi elementari così descritti:

$$\Omega = \{\underbrace{E_1, E_2, E_3}_{\textit{palline Verdi}}, \underbrace{E_4, E_5, E_6, E_7, E_8, E_9, E_{10}}_{\textit{palline Rosse}}\}$$

Allora, gli eventi V =«la pallina estratta è Verde» e R =«la pallina estratta è Rossa» coincideranno, rispettivamente, con gli eventi:

$$V = (E_1 \cup E_2 \cup E_3); \; R = (E_4 \cup E_5 \cup E_6 \cup E_7 \cup E_8 \cup E_9 \cup E_{10}).$$

Allora, la regola pratica può essere applicata perché il calcolo delle probabilità di tali eventi viene ricondotto a quello delle probabilità dei 10 eventi elementari E_i che costituiscono una partizione equiprobabile. Infatti, si ha:

$$Pr(V) = Pr(E_1 \cup E_2 \cup E_3) = Pr(E_1) + Pr(E_2) + Pr(E_3) = \frac{1}{10} + \frac{1}{10} + \frac{1}{10} = \frac{3}{10}.$$

In modo perfettamente analogo, per l'evento R =«la pallina estratta è Rossa» si ottiene:

$$\begin{aligned} Pr(R) &= Pr(E_4 \cup E_5 \cup \cdots \cup E_{10}) = Pr(E_4) + Pr(E_5) + \cdots + Pr(E_{10}) \\ &= \frac{1}{10} + \frac{1}{10} + \cdots + \frac{1}{10} = \frac{7}{10}. \end{aligned}$$

Come si vede, siamo riusciti a trasformare eventi non equiprobabili nella unione di eventi elementari equiprobabili, in modo da applicare a questi ultimi la precedente regola cui si è pervenuti per la misura della probabilità. Questo artifizio sarà sottinteso nel seguito, ed agevolerà di molto il calcolo delle probabilità in prove analoghe a queste.

• Nella prova «estrazione di 2 palline *con ripetizione*» da un'urna che contiene 4 palline Bianche e 6 Nere, si desidera calcolare la probabilità dell'evento E =«le due palline sono entrambe Bianche».

Ripetendo il precedente ragionamento, ciascuna pallina ha *singolarmente* probabilità 1/10 di essere estratta e, con ovvia simbologia, l'evento di interesse si può esprimere come: $E = (B_1 \cap B_2)$. Infatti, le 2 palline Bianche si verificheranno se e solo se si verificherà sia una pallina Bianca alla prima estrazione che una pallina Bianca alla seconda estrazione.

Ora, l'intersezione riguarda eventi indipendenti – perché l'estrazione avviene con ripetizione e quindi l'esito della prima sotto-prova non modifica la probabilità della seconda – per cui si ha:

$$P_r(E) = P_r(B_1 \cap B_2) = P_r(B_1)\ P_r(B_2) = \frac{4}{10}\frac{4}{10} = \frac{16}{100} = \frac{4}{25}.$$

Se, invece, la prova consiste nell'estrazione di 2 palline *senza ripetizione*, le probabilità di eventi connessi alla seconda sotto-prova si modificano in funzione degli esiti della prima sotto-prova. Pertanto, la medesima probabilità di 2 palline entrambe Bianche diventa:

$$P_r(E) = P_r(B_1 \cap B_2) = P_r(B_1)\ P_r(B_2|B_1) = \frac{4}{10}\frac{3}{9} = \frac{2}{15}.$$

Infatti, $P_r(B_2|B_1) = 3/9$ perché se la prima pallina estratta è Bianca allora nell'urna sono rimaste 3 palline Bianche su un totale di 9 palline.

• Nella prova «estrazione di due palline *senza ripetizione*» da un'urna che contiene 1 pallina Bianca e 9 Nere, il calcolo della probabilità che si verifichi l'evento E =«viene estratta una sola pallina Bianca» richiede una accurata formalizzazione e qualche precisazione.

Anzitutto, la prova consta di due *sottoprove non indipendenti* poiché la prima pallina estratta non viene rimessa nell'urna dopo la prima estrazione; pertanto, per calcolare la probabilità richiesta occorre formalizzare l'evento E tramite gli eventi elementari generati dalle due sottoprove.

Ora, in ciascuna estrazione, ogni singola pallina – a prescindere dal colore – ha probabilità costante e pari a 1/10 per la prima estrazione, e pari a 1/9 per la seconda (perché vi è una pallina in meno nell'urna). Pertanto, se indichiamo con B_1, B_2, N_1, N_2 gli eventi «pallina Bianca» e «pallina Nera», rispettivamente, alla prima e alla seconda estrazione, allora l'evento E =«viene estratta una sola pallina Bianca» si può esprimere tramite gli eventi generati dalle due sottoprove nel modo seguente: $E = (B_1 \cap N_2) \cup (N_1 \cap B_2)$. Infatti, questa espressione si legge: «E si verifica se si verifica Bianca nella prima estrazione e Nera nella seconda, oppure Nera nella prima estrazione e Bianca nella seconda». È anche immediato constatare che gli eventi $(B_1 \cap N_2)$ e $(N_1 \cap B_2)$ sono incompatibili.

Allora, applicando il Postulato 3 e ricordando le proprietà della intersezione di eventi non indipendenti, si avrà:

$$\begin{aligned} P_r(E) &= P_r((B_1 \cap N_2) \cup (N_1 \cap B_2)) = P_r(B_1 \cap N_2) + P_r(N_1 \cap B_2) \\ &= P_r(B_1)\,P_r(N_2|B_1) + P_r(N_1)\,P_r(B_2|N_1) = \frac{1}{10}\frac{9}{9} + \frac{9}{10}\frac{1}{9} = \frac{18}{90} = \frac{1}{5}. \end{aligned}$$

Negli ultimi passaggi, abbiamo così determinato le rispettive probabilità:

$P_r(B_1) = \frac{1}{10}$, perché nella 1^a estrazione vi è 1 pallina Bianca su 10 palline;

$P_r(N_2|B_1) = \frac{9}{9} = 1$, perché nella 2^a estrazione è certo che esca una pallina Nera se la 1^a pallina estratta è stata Bianca;

$P_r(N_1) = \frac{9}{10}$, perché nella 1^a estrazione vi sono 9 palline Nere su 10 palline;

$P_r(B_2|N_1) = \frac{1}{9}$, perché nella 2^a estrazione vi è 1 pallina Bianca su 9 se la 1^a pallina estratta è stata Nera.

• Ora, nella medesima prova, calcoliamo la probabilità dei seguenti eventi: A = «almeno una delle due palline è Bianca»; e C =«la seconda pallina estratta è Nera». Queste probabilità sono di calcolo agevole se si formalizzano accuratamente gli eventi elementari delle due sottoprove che li generano.

L'evento A si verifica se «la prima pallina estratta è Bianca e la seconda è Nera, oppure la prima estratta è Nera e la seconda è Bianca, oppure la prima è Bianca ed anche la seconda è Bianca». Quindi,

$$\begin{aligned} P_r(A) &= P_r((B_1 \cap N_2) \cup (N_1 \cap B_2) \cup (B_1 \cap B_2)) \\ &= P_r(B_1 \cap N_2) + P_r(N_1 \cap B_2) + P_r(B_1 \cap B_2) \\ &= P_r(B_1)\,P_r(N_2|B_1) + P_r(N_1)\,P_r(B_2|N_1) + P_r(B_1)\,P_r(B_2|B_1) \\ &= \frac{1}{10}\frac{9}{9} + \frac{9}{10}\frac{1}{9} + \frac{1}{10}0 = \frac{18}{90} = \frac{1}{5}. \end{aligned}$$

dove $P_r(B_2|B_1) = 0$, essendo impossibile che la seconda pallina sia Bianca se la prima pallina già estratta è risultata Bianca (infatti, la prova consiste nella estrazione senza ripetizione delle due palline e nell'urna vi è una sola pallina Bianca).

L'evento C si verifica se «la seconda pallina estratta è Nera». Tuttavia, occorre notare che la prova consiste comunque nell'estrazione di 2 palline, per cui non è possibile considerare questo evento senza tener conto di cosa avviene nella prima estrazione. Quindi, l'evento «la seconda pallina estratta è Nera» va correttamente descritto come «pallina Bianca nella prima estrazione e Nera nella seconda, oppure pallina Nera nella prima estrazione e Nera nella seconda». Formalmente, ciò implica:

$$\begin{aligned} P_r(C) &= P_r((B_1 \cap N_2) \cup (N_1 \cap N_2)) = P_r(B_1 \cap N_2) + P_r(N_1 \cap N_2) \\ &= P_r(B_1)\,P_r(N_2|B_1) + P_r(N_1)\,P_r(N_2|N_1) = \frac{1}{10}\frac{9}{9} + \frac{9}{10}\frac{8}{9} = \frac{81}{90} = \frac{9}{10}. \end{aligned}$$

dove: $P_r(N_2|N_1) = \frac{8}{9}$, perché se la prima pallina estratta è stata Nera, nella seconda sottoprova l'urna consisterà di 1 pallina Bianca e 8 palline Nere.

Si osservi che, in questa prova, lo spazio campione Ω può essere descritto come:

$$\Omega = \{(B_1 \cap B_2),\ (B_1 \cap N_2),\ (N_1 \cap B_2),\ (N_1 \cap N_2)\}$$

ed i quattro eventi elencati costituiscono una partizione di Ω perché necessari e incompatibili. Tuttavia, tali eventi non sono equiprobabili, perché:

$$\begin{aligned}
Pr(B_1 \cap B_2) &= Pr(B_1)\,Pr(B_2|B_1) = \frac{1}{10}\,0 = 0;\\
Pr(B_1 \cap N_2) &= Pr(B_1)\,Pr(N_2|B_1) = \frac{1}{10}\,1 = \frac{1}{10};\\
Pr(N_1 \cap B_2) &= Pr(N_1)\,Pr(B_2|N_1) = \frac{9}{10}\,\frac{1}{9} = \frac{1}{10};\\
Pr(N_1 \cap N_2) &= Pr(N_1)\,Pr(N_2|N_1) = \frac{9}{10}\,\frac{8}{9} = \frac{8}{10}.
\end{aligned}$$

Poiché i 4 eventi esaminati costituiscono una partizione di Ω, la somma di queste probabilità è, correttamente, pari a 1 (per il secondo Postulato).

□

Anche se l'ipotesi di equiprobabilità è semplice e, talora, convincente, non sempre è possibile invocarla. Tuttavia, in alcune situazioni, è ancora possibile far derivare l'assegnazione della probabilità a partire dai Postulati, come mostrano i casi illustrati nell'esempio seguente.

Esempio 8.9

• Un dado è stato truccato in modo che la probabilità di ogni faccia sia proporzionale al punteggio che vi è impresso. Così, l'evento E_6 è 6 volte più probabile dell'evento E_1 e 3 volte più probabile dell'evento E_2, e così via, per tutti gli eventi elementari.

Anche per un dado del genere valgono i Postulati del Calcolo delle probabilità e tramite essi è possibile assegnare la probabilità a ciascun evento elementare in modo univoco.

Infatti, se poniamo: $Pr(E_1) = p$ (dove $p \geq 0$ in ossequio al Postulato 1), allora per i Postulati 2 e 3, tenuto conto che il presentarsi di ciascuna delle 6 facce rappresenta una collezione di eventi incompatibili, deve essere:

$$\begin{aligned}
1 &= Pr(\Omega) = Pr(E_1 \cup E_2 \cup \ldots E_6) = Pr(E_1) + Pr(E_2) + \cdots + Pr(E_6)\\
&= p + 2p + \cdots + 6p = 21\,p
\end{aligned}$$

il che implica necessariamente: $p = \frac{1}{21}$ e, di conseguenza,

$$Pr(E_1) = \frac{1}{21};\quad Pr(E_2) = \frac{2}{21};\quad Pr(E_3) = \frac{3}{21};$$

$$Pr(E_4) = \frac{4}{21};\quad Pr(E_5) = \frac{5}{21};\quad Pr(E_6) = \frac{6}{21}.$$

A tal punto, avendo determinata la probabilità degli eventi elementari di Ω, è possibile determinare coerentemente la probabilità di qualsiasi evento deducibile dagli eventi elementari mediante un numero finito o numerabile di unioni, intersezioni e negazioni. Per

esempio, in una prova del genere, la probabilità che si verifichi una faccia pari è sensibilmente superiore a quella di una faccia dispari (è intuitivo: ogni numero dispari ha una probabilità inferiore a quello successivo). Formalmente, si ha:

$$\begin{aligned}
Pr(\text{faccia dispari}) &= Pr(E_1) + Pr(E_3) + Pr(E_5) = \frac{1}{21} + \frac{3}{21} + \frac{5}{21} = \frac{9}{21} = 0.42857\\
Pr(\text{faccia pari}) &= Pr(E_2) + Pr(E_4) + Pr(E_6) = \frac{2}{21} + \frac{4}{21} + \frac{6}{21} = \frac{12}{21} = 0.57143
\end{aligned}$$

Ancora, si possono calcolare altre probabilità. Per esempio, quella che si verifichi la faccia con 2 punti essendosi verificato un numero pari (per un dato corretto tale probabilità vale 1/3, ovviamente):

$$\begin{aligned}
Pr(E_2|(E_2 \cup E_4 \cup E_6)) &= \frac{Pr(E_2 \cap (E_2 \cup E_4 \cup E_6))}{Pr(E_2 \cup E_4 \cup E_6)} = \frac{Pr(E_2)}{Pr(E_2 \cup E_4 \cup E_6)}\\
&= \frac{2/21}{12/21} = \frac{2}{12} = \frac{1}{6} = 0.16667
\end{aligned}$$

• Nelle scommesse (sui cavalli, sulle partite di calcio, sulle corse, etc.) si parla spesso di «quotazioni» (=*odds*) per un evento E anziché della sua probabilità, perché si ritiene che ciò sia più immediato per una valutazione comparata delle possibilità di vincere o di perdere scommettendo su E. Tuttavia, anche questa valutazione conduce ad una determinazione della probabilità coerente con i Postulati.

Per esempio, dire che la vittoria (evento E) di una squadra è quotata 3 a 1 (oppure 3 : 1) significa affermare che si ritiene 3 volte più probabile che si verifichi l'evento E piuttosto che l'evento $\overline{E}$. È agevole constatare che questa è una assegnazione corretta della probabilità perché:

$$\frac{3}{1} = \frac{Pr(E)}{Pr(\overline{E})} = \frac{Pr(E)}{1 - Pr(E)} \Longrightarrow Pr(E) = \frac{3}{4}.$$

In generale, assegnare ad un evento E una quotazione pari a $m : n$ equivale ad assegnare ad E la probabilità: $Pr(E) = \dfrac{m}{m+n}$.

□

L'assegnazione della probabilità è possibile anche quando lo spazio campione è costituito da un'infinità numerabile di eventi elementari, come mostra l'esempio seguente, catalogabile nella tipologia di prove «attesa dell'evento» (discusse nel paragrafo 8.7).

Esempio 8.10

• Supponiamo che una moneta possieda probabilità $\theta \in (0,1)$ che si verifichi «Testa». La prova consiste nel lanciare ripetutamente tale moneta fino a che esca «Testa» per la prima volta.

Tale prova è caratterizzata da uno spazio campione: $\Omega = \{E_1, E_2, \ldots\}$ i cui eventi elementari – con ovvia simbologia – sono così elencabili:

$E_1 = T_1$ (si verifica subito Testa: la prova termina al 1° lancio);

$E_2 = C_1 \cap T_2$ (si verifica Croce e poi Testa: la prova termina al 2° lancio);

$E_3 = C_1 \cap C_2 \cap T_3$ (si verifica 2 volte Croce e poi Testa: la prova termina al 3° lancio);

$\vdots$

$E_i = C_1 \cap C_2 \cap \cdots \cap C_{i-1} \cap T_i$ (si verifica $i-1$ volte Croce e poi Testa: la prova termina al lancio i-esimo);

$\vdots$

Poiché, in ciascun lancio, sono costanti sia $Pr(T) = \theta$ che $Pr(C) = 1 - \theta$, tenendo presente che i lanci successivi sono una successione di sottoprove indipendenti, agli eventi precedentemente elencati si possono allora assegnare le seguenti probabilità:

$Pr(E_1) = Pr(T_1) = \theta$;
$Pr(E_2) = Pr(C_1 \cap T_2) = (1-\theta)\,\theta$;
$Pr(E_3) = Pr(C_1 \cap C_2 \cap T_3) = (1-\theta)\,(1-\theta)\,\theta = (1-\theta)^2\,\theta$;

$\vdots$

$Pr(E_i) = Pr(C_1 \cap C_2 \cap \cdots \cap C_{i-1} \cap T_i) = (1-\theta)\,(1-\theta)\,\ldots\,\theta = (1-\theta)^{i-1}\,\theta$;

$\vdots$

Tale assegnazione è ben definita perché rispetta i Postulati; in particolare, ciascuna probabilità è non-negativa ed inoltre (ricordando la somma di una serie geometrica):

$$Pr(\Omega) = \sum_{i=1}^{\infty} Pr(E_i) = \sum_{i=1}^{\infty} (1-\theta)^{i-1}\theta = \theta \sum_{i=0}^{\infty} (1-\theta)^i = \theta\,\frac{1}{1-(1-\theta)} = 1.$$

Possiamo, allora, legittimamente calcolare su tale prova tutte le probabilità che ci interessano. Per esempio, la probabilità che la prova termini entro il 3° lancio (cioè al 1°, al 2° oppure al 3°) è pari a:

$$Pr(E_1 \cup E_2 \cup E_3) = Pr(E_1) + Pr(E_2) + Pr(E_3) = \theta + (1-\theta)\,\theta + (1-\theta)^2\,\theta = 1 - (1-\theta)^3.$$

La probabilità che la prova si concluda dopo un numero pari di lanci, cioè la probabilità dell'evento $(E_2 \cup E_4 \cup \cdots \cup E_{2i} \cup \ldots)$, dopo alcuni passaggi algebrici, è:

$$\begin{aligned} \sum_{i=1}^{\infty} Pr(E_{2i}) &= \sum_{i=1}^{\infty} (1-\theta)^{2i-1}\,\theta = \theta\,(1-\theta) \sum_{i=0}^{\infty} \left[(1-\theta)^2\right]^i \\ &= \theta\,(1-\theta)\,\frac{1}{1-(1-\theta)^2} = \frac{1-\theta}{2-\theta}. \end{aligned}$$

Evidentemente, la probabilità che la prova si concluda dopo un numero dispari di lanci, cioè la probabilità dell'evento $(E_1 \cup E_3 \cup \cdots \cup E_{2i-1} \cup \ldots)$, sarà uguale a:

$$\sum_{i=1}^{\infty} Pr(E_{2i-1}) = 1 - \sum_{i=1}^{\infty} Pr(E_{2i}) = 1 - \frac{1-\theta}{2-\theta} = \frac{1}{2-\theta}.$$

Ebbene, è agevole dimostrare che la probabilità che la prova termini dopo un numero dispari di lanci è sempre superiore alla probabilità che la prova termini dopo un numero pari

di lanci, per quanto truccata sia la moneta (cioè qualsiasi sia il valore di θ). In particolare, se la moneta è corretta, cioè se $\theta = 1/2$, tali probabilità valgono rispettivamente:

$$\begin{aligned} \sum_{i=1}^{\infty} Pr(E_{2i}) &= \frac{1-1/2}{2-1/2} = \frac{1}{3} = 0.33333; \\ \sum_{i=1}^{\infty} Pr(E_{2i-1}) &= \frac{1}{2-2-1/2} = \frac{2}{3} = 0.66667. \end{aligned}$$

Il fatto che la probabilità che la prova termini dopo un numero dispari di lanci sia sempre superiore alla probabilità che la prova termini dopo un numero pari di lanci era derivabile anche su base intuitiva. Infatti, per qualunque valore di θ, si ha:

$$Pr(E_{2i-1}) > Pr(E_{2i}), \forall i = 1, 2, \ldots,$$

per cui sommando membro a membro si deriva il risultato.

• Infine, va sottolineato come la prova qui schematizzata non debba necessariamente essere interpretata con riferimento al lancio ripetuto di una moneta. Infatti, essa si applica a qualsiasi prova che consiste nella replica di sottoprove indipendenti nelle quali sia costante e pari a θ la probabilità di un evento E, di nostro interesse, e le sottoprove continuano sino a che non si verifica E per la prima volta.

Per esempio, se si lanciano 2 dadi ben equilibrati sino a che non esce un «doppio 6» siamo in una prova dello stesso tipo con $\theta = 1/36$, per cui – applicando i risultati precedenti – si deduce che tale prova termina dopo un numero pari di lanci con probabilità $35/71 = 0.49296$. Ancora, se ciascuno di un gruppo di amici estrae *con ripetizione*, a turno, una carta da un mazzo di 40 carte napoletane sino a che qualcuno non estrae uno dei quattro assi, la prova è ancora del medesimo tipo con $\theta = 1/10$ per cui, per esempio, tale gioco termina dopo un numero dispari di estrazioni con probabilità $10/19 = 0.52632$; e così via. □

Gli esempi hanno confermato che *qualunque* assegnazione di probabilità che attribuisca agli eventi elementari E_i una successione di numeri non negativi $p_i \geq 0$ e tali che la loro somma sia 1 è coerente con i Postulati del Calcolo delle probabilità.

Il problema si complica notevolmente se gli eventi elementari non sono in numero finito o infinito numerabile (cioè se non sono elencabili), come avviene se i possibili risultati della prova sono in corrispondenza con tutti i punti di un qualsiasi intervallo reale.

o Esempi di prove che generano risultati i cui esiti sono *concettualmente* assimilabili ai punti di un intervallo continuo sono: l'altezza di una persona adulta scelta a caso; la durata di funzionamento di una lampadina; la distanza a cui ricade un peso scagliato il più lontano possibile; la frazione di ossido di carbonio rilevata a mezzogiorno al centro di una città; il tempo che intercorre tra la fine di una conversazione telefonica e l'inizio della successiva per un certo numero di cellulare, etc.

Limitandoci al primo esempio, se ipotizziamo di aver trovato due persone alte rispettivamente mm 1680 e 1681 nessuno può escludere che esista una persona la cui altezza assume qualsiasi valore interno all'intervallo (1680, 1681). In effetti, l'altezza può *concettualmente* assumere tutti i valori di un intervallo congruo, per esempio (1350, 2100).

In breve, la difficoltà di assegnare misure di probabilità a spazi non-numerabili deriva dalla impossibilità di porre in corrispondenza gli eventi elementari che hanno la potenza del continuo con l'insieme dei numeri naturali, la cui potenza è quella di un insieme numerabile. In altri termini, non è possibile «elencare le probabilità dei singoli eventi garantendo che esse siano tutte non-negative e contemporaneamente sommino ad 1».

Il problema viene superato mediante una procedura matematica che esula dal livello che ci siamo imposti in questo volume.

8.13 Il Teorema di Bayes

Il Teorema di Bayes è una applicazione molto semplice del teorema delle probabilità totali (paragrafo 8.11) ma ha assunto un ruolo di notevole importanza per la interpretazione statistica che se ne può derivare.

- Tale interpretazione è divenuta così estesa che una impostazione recente della Statistica viene definita «Statistica Bayesiana», in contrapposizione a quella più consolidata, definita «Statistica classica».
- Il teorema è attribuito al rev. Thomas Bayes e fu reso noto dal suo amico Price che, tre anni dopo la sua morte, trovò tra le sue carte il manoscritto e lo pubblicò nel 1764. Esso è stato utilizzato soprattutto da Laplace.
- Nonostante sia tuttora aperta la discussione circa la sua effettiva utilizzabilità nei casi reali, occorre riconoscere che tale teorema costituisce una pietra miliare nella teoria della conoscenza statistica perché esplicita il ruolo delle differenti probabilità nella scelta tra decisioni alternative.

▶ ***TEOREMA DI BAYES*** Se $H_1, H_2, \ldots, H_m$ sono eventi che costituiscono una partizione di Ω, allora per qualsiasi evento $E \subset \Omega$ si ha:

$$P_r(H_i|E) = \frac{P_r(H_i)\ P_r(E|H_i)}{\sum_{j=1}^{m} P_r(H_j)\ P_r(E|H_j)}, \quad \forall i = 1, 2, \ldots, m.$$

Dimostrazione. Per ogni evento $E_i \subset \Omega$, si ha:

$$P_r(H_i|E) = \frac{P_r(H_i \cap E)}{P_r(E)} = \frac{P_r(H_i)\, P_r(E|H_i)}{P_r(E)}, \forall i = 1, 2, \ldots, m.$$

Ora, essendo gli eventi $H_j, j = 1, 2, \ldots, m$ una partizione di Ω, per il teorema delle probabilità totali (Teorema 8.9), la probabilità di E è espressa da:

$$P_r(E) = \sum_{j=1}^{m} P_r(H_j)\ P_r(E|H_j).$$

E, sostituendo tale espressione al denominatore della formula precedente, si ottiene il Teorema di Bayes.

c.v.d.

- Si osservi che la somma delle probabilità $P_r(H_i|E)$ per $i = 1, 2, \ldots, m$ vale esattamente 1 perché gli eventi $H_1, H_2, \ldots, H_m$ sono una partizione di Ω.

Inoltre, la dimostrazione del teorema continua a sussistere anche se m è infinito, cioè se Ω è costituito da una successione (numerabile) di eventi elementari $H_1, H_2, \ldots$ necessari ed incompatibili.

Nella sua essenza, il Teorema di Bayes mostra come la probabilità dell'affermazione «E si è verificato» modifica l'assegnazione delle probabilità $P_r(H_i)$ nelle probabilità condizionate $P_r(H_i|E)$. Ovvero, il Teorema di Bayes esplicita le probabilità condizionate $P_r(H_i|E)$ tramite le probabilità $P_r(E|H_i)$ e $P_r(H_i)$.

L'utilizzo del Teorema di Bayes si estende in modo notevole attribuendo un particolare significato agli eventi che sono presenti nella sua espressione formale.

In particolare, supponiamo che E sia un evento (=*effetto*) che possa realizzarsi in conseguenza di m «cause» H_i, una delle quali certamente agisce e ognuna delle quali ha una *probabilità a priori* $P_r(H_i)$ di agire. Supponiamo anche che $P_r(E|H_i)$ sia la probabilità che E si verifichi quando è noto che abbia agito la causa H_i, e chiameremo tale probabilità la *verosimiglianza* di E dato H_i.

Allora, dato un effetto E ed m cause H_i, necessarie ed incompatibili, *il Teorema di Bayes esprime la* ***probabilità a posteriori*** $P_r(H_i|E)$, *cioè la probabilità che avendo osservato l'evento E esso sia stato generato dalla causa* H_i, *in funzione delle* ***probabilità a priori*** $P_r(H_i)$ *e delle* ***verosimiglianze*** $P_r(E|H_i)$.

Questa interpretazione rende il Teorema di Bayes un momento fondamentale della ricerca scientifica poiché di fronte ad un risultato osservato, un effetto noto, un dato sperimentale (l'evento E), che può essere stato generato in associazione e/o in dipendenza di più cause, teorie, ipotesi, leggi (gli eventi necessari ed incompatibili H_i), è possibile calcolare la probabilità che abbiano agito le differenti cause H_i, e quindi offrire a colui che deve assumere una decisione un criterio di scelta molto ragionevole (per esempio, suggerendogli di optare per quella «causa» H_i che rende massima la probabilità a posteriori $P_r(H_i|E)$).

- Poiché il denominatore presente nella formula del Teorema di Bayes, cioè la probabilità di E non varia per $i = 1, 2, \ldots$, il suo ruolo è quello di fungere da costante di normalizzazione, per far sì che la somma delle probabilità a posteriori sia pari a 1. Pertanto, per ogni $i = 1, 2, \ldots, m$, si può anche scrivere:

$$P_r(H_i|E) \propto P_r(H_i) \times P_r(E|H_i)$$

ovvero

$$\textit{probabilità a posteriori} \propto (\textit{probabilità a priori}) \times (\textit{verosimiglianza}).$$

- La discussione precedente spiega perché il Teorema di Bayes interviene nel cosiddetto *problema inverso*, cioè nel problema di determinare le cause più probabili avendo osservato solo i risultati sperimentali.

Va aggiunto, però, che a differenza degli esempi successivi (nei quali tutte le probabilità sono agevolmente deducibili dalla struttura della prova oppure vengono supposte note), è piuttosto raro pervenire a valutazioni incontrovertibili delle probabilità a priori. Il che

provoca interminabili discussioni circa la effettiva applicazione del Teorema di Bayes nella soluzione di casi reali.

Esempio 8.11

Supponiamo che un medico sappia che un certo sintomo E (per esempio, una febbre altissima associata ad un quadro clinico specifico) possa essere l'effetto di tre sole malattie H_1, H_2, H_3 con probabilità $Pr(E|H_1), Pr(E|H_2), Pr(E|H_3)$ che sono le rispettive *verosimiglianze*; esse sono il risultato della ricerca medica e sono generalmente contenute nei trattati di Medicina.

Peraltro, la sua esperienza di medico, lo studio del territorio e delle sue condizioni igienico-sanitarie nonché le opinioni personali che si è formato negli anni lo inducono a credere che tali malattie siano presenti con probabilità $Pr(H_1), Pr(H_2), Pr(H_3)$, che sono le *probabilità a priori*, così definite perché non sono determinate con riferimento all'evento E. Esse non derivano da una valutazione riferita ad uno specifico paziente, ma sono una quantificazione delle opinioni del medico circa la possibilità che agiscano le tre malattie.

Quando, poi, nel suo studio si presenta un paziente specifico con il sintomo E, il medico deve decidere quale delle tre malattie abbia probabilmente agito, poiché la terapia che prescriverà varia in funzione della presenza (solo probabile) di una fra le tre malattie H_1, H_2, H_3. In buona sostanza, egli deve calcolare $Pr(H_1|E), Pr(H_2|E), Pr(H_3|E)$, che sono le *probabilità a posteriori* che abbia agito una delle tre malattie avendo riscontrato il sintomo E.

Nel pervenire ad una decisione, il medico determina la malattia più probabile (prescrivendo, quindi, la relativa terapia) applicando il Teorema di Bayes cioè contemperando – in senso moltiplicativo – i valori delle probabilità a priori con le rispettive verosimiglianze e confrontando, alla fine, le probabilità a posteriori. In effetti, la ricerca medica è ben più complessa di questo esempio perché, di fronte ad ogni paziente, l'anamnesi aiuta (o dovrebbe aiutare) nella migliore formulazione delle probabilità a posteriori; per questo, è generalmente possibile rendere più realistica la situazione qui illustrata.

Per fornire una valutazione numerica, supponiamo che il medico stabilisca le sue probabilità a priori e le verosimiglianze nel modo seguente:

$$\begin{array}{llllll} Pr(H_1) = 0.03; & Pr(H_2) = 0.70; & Pr(H_3) = 0.27; \\ Pr(E|H_1) = 0.90; & Pr(E|H_2) = 0.10; & Pr(E|H_3) = 0.30. \end{array}$$

Si nota che, *a priori*, la malattia H_2 è la più probabile, ma il sintomo E è con maggiore verosimiglianza associato ad H_1 la quale, di per sé, sembra essere a priori una malattia poco diffusa (infatti, la sua probabilità a priori è uguale a 0.03).

Per applicare il Teorema di Bayes, è conveniente calcolare il denominatore della formula che, nello specifico, è pari a:

$$\begin{aligned} Pr(E) &= Pr(H_1)\,Pr(E|H_1) + Pr(H_2)\,Pr(E|H_2) + Pr(H_3)\,Pr(E|H_3) \\ &= (0.03)(0.90) + (0.70)(0.10) + (0.27)(0.30) = 0.178 \end{aligned}$$

Allora, le tre probabilità *a posteriori* risultano essere:

$$\begin{aligned} Pr(H_1|E) &= \frac{Pr(H_1)\,Pr(E|H_1)}{Pr(H_1)\,Pr(E|H_1) + Pr(H_2)\,Pr(E|H_2) + Pr(H_3)\,Pr(E|H_3)} \\ &= \frac{(0.03)(0.90)}{(0.03)(0.90) + (0.70)(0.10) + (0.27)(0.30)} = 0.15169 \end{aligned}$$

$$\begin{aligned} Pr(H_2|E) &= \frac{Pr(H_2)\,Pr(E|H_2)}{Pr(H_1)\,Pr(E|H_1) + Pr(H_2)\,Pr(E|H_2) + Pr(H_3)\,Pr(E|H_3)} \\ &= \frac{(0.70)(0.10)}{(0.03)(0.90) + (0.70)(0.10) + (0.27)(0.30)} = 0.39326 \end{aligned}$$

$$\begin{aligned} Pr(H_3|E) &= \frac{Pr(H_3)\,Pr(E|H_3)}{Pr(H_1)\,Pr(E|H_1) + Pr(H_2)\,Pr(E|H_2) + Pr(H_3)\,Pr(E|H_3)} \\ &= \frac{(0.27)(0.30)}{(0.03)(0.90) + (0.70)(0.10) + (0.27)(0.30)} = 0.45505 \end{aligned}$$

Come si vede, la presenza del sintomo E ha modificato l'opinione del medico circa la graduatoria delle malattie; infatti, *a priori* la malattia più probabile era H_2, seguita da H_3 e, infine, da H_1. Per contro, la presenza di E ha fatto sì che il medico debba ora propendere, *a posteriori*, anzitutto per la malattia H_3, poi per H_2 e, infine, per H_1. □

L'esempio ha mostrato come la graduatoria tra le probabilità a posteriori sia determinata solo dal numeratore della formula del Teorema di Bayes. Ciò è ancora più evidente nel caso di due sole «cause» o «ipotesi», H_0 e H_1. Si propenderà per H_1 (rispetto ad H_0) se la probabilità a posteriori di H_1 dato E sarà maggiore della probabilità a posteriori di H_0 dato E, cioè se:

$$\begin{aligned} Pr(H_1|E) > Pr(H_0|E) &\Rightarrow \frac{Pr(H_1)\;Pr(E|H_1)}{\sum_{j=0}^{1} Pr(H_j)\;Pr(E|H_j)} > \frac{Pr(H_0)\;Pr(E|H_0)}{\sum_{j=0}^{1} Pr(H_j)\;Pr(E|H_j)} \\ &\Rightarrow Pr(H_1)\;Pr(E|H_1) > Pr(H_0)\;Pr(E|H_0) \\ &\Rightarrow \left(\frac{Pr(H_1)}{Pr(H_0)}\right)\left(\frac{Pr(E|H_1)}{Pr(E|H_0)}\right) > 1. \end{aligned}$$

L'ultima espressione mostra che il rapporto tra le probabilità a posteriori dipende dal prodotto dei due rapporti indicati nelle parentesi dell'ultimo passaggio: il *rapporto tra le probabilità a priori* (detto anche «a priori odds») e il *rapporto tra le verosimiglianze* (detto anche «fattore di Bayes»). Entrambi contribuiscono a far sì che il rapporto tra le probabilità a posteriori possa essere superiore all'unità (il che fa propendere la scelta per H_1) oppure inferiore all'unità (il che fa propendere la scelta per H_0).

Ora, se l'evidenza dei fatti è bassa per favorire un'ipotesi anziché l'altra (il che implica che il rapporto tra le verosimiglianze sia circa 1, cioè che entrambe le ipotesi potevano dar luogo all'evento con probabilità molto simili), la decisione si basa sul rapporto tra le probabilità a priori. È interessante constatare come piccole deviazioni rispetto alla indifferenza possano contribuire a modificare radicalmente il rapporto tra le probabilità a posteriori.

o Se $Pr(H_0) \simeq Pr(H_1)$, allora $Pr(H_1)/Pr(H_0) \simeq 1$, e quindi la decisione dipende esclusivamente dal rapporto tra le verosimiglianze: $Pr(E|H_1)/Pr(E|H_0)$.

Se, invece, $Pr(H_1) = 0.55, Pr(H_0) = 0.45$ (quindi una ipotesi è leggermente preferibile all'altra), allora $Pr(H_1)/Pr(H_0) = 1.22$, e quindi l'evidenza a priori apporta al rapporto tra le verosimiglianze un contributo che è superiore del 22% mentre la differenza tra le probabilità a priori è di appena 0.10.

Se $Pr(H_1) = 0.60, Pr(H_0) = 0.40$, allora $Pr(H_1)/Pr(H_0) = 1.50$, e quindi l'evidenza a priori aumenta il rapporto tra le verosimiglianze del 50%; e così via.

Questi esempi numerici mostrano che anche modesti allontanamenti dall'ipotesi di indifferenza *a priori* determinano un'alterazione elevata, più di quanto si creda, alla scelta *a posteriori* tra le differenti ipotesi. In altri termini, l'evidenza dei fatti può essere notevolmente alterata da convincimenti a priori circa le ipotesi.

Le conseguenze del Teorema di Bayes per stabilire il valore dell'evidenza dedotta da prove fornite o accertate nel processo penale sono discusse con ampiezza nella letteratura. In questi casi reali si evidenzia correttamente il ruolo dello statistico, che non è quello di formulare giudizi ma di fornire valutazioni probabilistiche circa l'evidenza delle situazioni in rapporto alle differenti ipotesi. Infatti, è compito del giudice sintetizzare, assieme alle altre, queste valutazioni probabilistiche e formulare un giudizio finale.

Esempio 8.12

Una prova consiste nel lanciare un dado ben equilibrato e nell'estrarre poi, a seconda del risultato del dado, una pallina da una di tre urne U_1, U_2, U_3 (contenenti palline Bianche e Nere in proporzioni differenti).

Indicheremo le 6 facce del dado con $D_i, i = 1, 2, \ldots, 6$ mentre le tre urne hanno la seguente composizione:

$$U_1 \equiv \{9\ Bianche,\ 1\ Nera\};\ U_2 \equiv \{1\ Bianca,\ 9\ Nere\};\ U_3 \equiv \{5\ Bianche,\ 5\ Nere\}.$$

La prova consiste nella procedura seguente:

se si verifica D_1,	si estrae 1 pallina dall'urna U_1;
se si verifica $D_2 \cup D_3 \cup D_4$,	si estrae 1 pallina dall'urna U_2;
se si verifica $D_5 \cup D_6$,	si estrae 1 pallina dall'urna U_3.

La probabilità che, alla fine della prova, si generi l'evento B =«pallina Bianca» si ottiene tramite il teorema delle probabilità totali, tenendo presente che la pallina Bianca può essere estratta da ciascuna delle tre urne. Con evidente simbologia, si ha:

$$\begin{aligned} Pr(B) &= Pr((U_1 \cap B_1) \cup (U_2 \cap B_2) \cup (U_3 \cap B_3)) \\ &= Pr(U_1)\,Pr(B_1|U_1) + Pr(U_2)\,Pr(B_2|U_2) + Pr(U_3)\,Pr(B_3|U_3) \\ &= \frac{1}{6}\frac{9}{10} + \frac{3}{6}\frac{1}{10} + \frac{2}{6}\frac{5}{10} = \frac{22}{60} = \frac{11}{30} = 0.36667 \end{aligned}$$

Tale probabilità costituisce la soluzione di un *problema diretto*. Invece, il Teorema di Bayes risponde al seguente *problema inverso*: supponiamo che si estragga una pallina e che essa risulti Bianca. Qual è la probabilità che essa sia stata estratta dalla 1^a urna, oppure dalla 2^a, oppure dalla 3^a?

Le probabilità di questi eventi si calcolano utilizzando il Teorema di Bayes, cioè nel modo seguente:

$$\begin{aligned} Pr(U_1|B) &= \frac{Pr(U_1)\,Pr(B_1|U_1)}{Pr(U_1)\,Pr(B_1|U_1) + Pr(U_2)\,Pr(B_2|U_2) + Pr(U_3)\,Pr(B_3|U_3)} \\ &= \frac{(1/6)(9/10)}{(1/6)(9/10) + (3/6)(1/10) + (2/6)(5/10)} = \frac{9}{22} = 0.40909 \end{aligned}$$

$$\begin{aligned} Pr(U_2|B) &= \frac{Pr(U_2)\,Pr(B_2|U_2)}{Pr(U_1)\,Pr(B_1|U_1) + Pr(U_2)\,Pr(B_2|U_2) + Pr(U_3)\,Pr(B_3|U_3)} \\ &= \frac{(3/6)(1/10)}{(1/6)(9/10) + (3/6)(1/10) + (2/6)(5/10)} = \frac{3}{22} = 0.13636 \end{aligned}$$

$$\begin{aligned} Pr(U_3|B) &= \frac{Pr(U_3)\,Pr(B_3|U_3)}{Pr(U_1)\,Pr(B_1|U_1) + Pr(U_2)\,Pr(B_2|U_2) + Pr(U_3)\,Pr(B_3|U_3)} \\ &= \frac{(2/6)(5/10)}{(1/6)(9/10) + (3/6)(1/10) + (2/6)(5/10)} = \frac{10}{22} = 0.45455 \end{aligned}$$

È interessante confrontare le *probabilità a priori* con le *probabilità a posteriori* delle tre urne:

$Pr(U_1)$=0.16667;	$Pr(U_1\|B)$=0.40909
$Pr(U_2)$=0.50000;	$Pr(U_2\|B)$=0.13636
$Pr(U_3)$=0.33333;	$Pr(U_3\|B)$=0.45455

Quindi, il presentarsi di una pallina Bianca ha modificato in modo sostanziale le probabilità di estrarre una pallina bianca dalle tre differenti urne.

□

Esempio 8.13

Questo esempio è ancora di interesse medico e riguarda il rischio di analisi di laboratorio errate. Supponiamo che un virus piuttosto raro sia presente nella popolazione con frequenza 0.00001 cioè solo 1 persona su 100.000 è portatore di quel virus. Un test clinico è stato sperimentato estesamente e risulta accertato che il 98% delle volte diagnostica correttamente la presenza di quel virus, ma per l'1% delle volte diagnostica il virus anche per una persona perfettamente immune (il cosiddetto *falso positivo*).

Il problema che ci si pone è: *se una persona si sottopone al test e questo risulta positivo, qual è la probabilità che la persona effettivamente abbia quel virus?*

Definiamo i seguenti eventi:

T_1	=	«il test diagnostica la presenza del virus»;
T_0	=	«il test non diagnostica la presenza del virus»;
V_1	=	«la persona ha effettivamente il virus»;
V_0	=	«la persona non ha effettivamente il virus».

Per rispondere alla precedente domanda, evidentemente, occorre calcolare la probabilità $Pr(V_1|T_1)$ conoscendo le probabilità a priori:

$$Pr(V_1) = 0.00001; \quad Pr(V_0) = 0.99999;$$

e le verosimiglianze:

$$Pr(T_1|V_1) = 0.98; \quad Pr(T_1|V_0) = 0.01.$$

Applicando il teorema di Bayes si ha:

$$\begin{aligned} Pr(V_1|T_1) &= \frac{Pr(V_1)\,Pr(T_1|V_1)}{Pr(V_0)\,Pr(T_1|V_0) + Pr(V_1)\,Pr(T_1|V_1)} \\ &= \frac{(0.00001)(0.98)}{(0.99999)(0.01) + (0.00001)(0.98)} = 0.00098 \end{aligned}$$

Quindi, se il test diagnostica la presenza del virus vi è una probabilità inferiore all'1 per mille che effettivamente la persona abbia quel virus. Tale risultato, in sostanza, dipende dalla estrema rarità del virus nella popolazione in esame.

□

8.14 Alcune esemplificazioni

In questo paragrafo presentiamo alcune applicazioni che, per ragioni storiche e didattiche, sono spesso riportate nei testi di Calcolo delle probabilità. In effetti, in molte situazioni sperimentali così come in esercitazioni appositamente predisposte, il calcolo della probabilità di un evento E richiede di enumerare correttamente gli eventi elementari «favorevoli» ad E da rapportare al numero di quelli complessivi dello spazio Ω. Poiché non possiamo presupporre nei lettori adeguate conoscenze di analisi combinatoria, abbiamo escluso dagli esempi tale tipo di prove (per un esempio di questa tipologia, si veda il paragrafo R 8.3).

Invece, riteniamo utile premettere uno *schema operativo* che semplifica il calcolo della probabilità di un evento E, anche complesso, perché lo riconduce ad una serie di operazioni semplici su eventi elementari e mediante un percorso ben definito. In sostanza, prima di affrontare direttamente la misura della probabilità di un evento, in qualsiasi contesto, noi suggeriamo di svolgere tutte le tappe dello schema seguente.

1. Individuare correttamente la *prova* e tutti gli *eventi elementari* di cui è composta, distinguendo tra prove e sottoprove.

2. Assegnare la probabilità agli eventi elementari seguendo la formalizzazione dei paragrafi precedenti e, quindi, distinguendo tra:

   i) partizioni finite di eventi;

   ii) partizioni numerabili di eventi;

   iii) partizioni non-numerabili di eventi.

   Dedurre, poi, a seconda dei casi, la misura della probabilità degli eventi elementari da ragioni di simmetria, dalla struttura fisica dell'esperimento, da argomentazioni geometriche o analitiche, da esperienze precedenti, da valutazioni soggettive.

3. Controllare sempre che l'assegnazione delle probabilità agli eventi elementari rispetti i Postulati, cioè le condizioni necessarie e sufficienti perché essa sia coerente.

4. Esplicitare l'evento E di cui si desidera calcolare la probabilità mediante la unione, la negazione, la intersezione degli eventi elementari, in numero finito o numerabile. Ciò è sempre possibile per qualsiasi evento generato dalla prova.

5. Possono, allora, verificarsi solo tre casi:

   i) se l'evento E è costituito dall'*unione* di eventi elementari, occorre chiedersi se essi sono incompatibili o meno, applicando poi le rispettive formule;

   ii) se l'evento E è costituito dalla *intersezione* di eventi elementari, occorre chiedersi se essi sono indipendenti o meno, applicando poi le rispettive formule;

   iii) se l'evento E è rappresentato dalla *negazione* di eventi elementari (o dalla loro unione o dalla loro intersezione), occorre applicare anzitutto le formule delle probabilità per l'evento negazione, ovvero le leggi di de Morgan, se esse risultano convenienti.

6. Qualora l'evento E derivi da più sottoprove, occorre verificarne la indipendenza oppure calcolare le probabilità condizionate corrispondenti ai singoli eventi, applicando il teorema delle probabilità totali oppure il Teorema di Bayes.

7. Controllare, alla fine di ogni elaborazione numerica, la coerenza dei risultati ottenuti, per esempio, calcolando in modo differente la probabilità dell'evento E oppure di altri eventi collegati ad E.

Negli esempi seguenti, con gli opportuni adattamenti, cercheremo di utilizzare sempre tale schema.

Esempio 8.14

Supponiamo che la prova consista nella *estrazione senza ripetizione* di 2 palline da un'urna che contiene 3 palline Bianche e 7 Nere. Si desidera calcolare la probabilità dei seguenti eventi:

A ={la prima pallina è Nera e la seconda è Bianca};
C ={la prima pallina possiede colore differente dalla seconda};
D ={le due palline estratte hanno lo stesso colore};
E ={almeno una tra le due palline estratte è Bianca};
F ={la seconda pallina estratta è Bianca}.

Ora, estraendo due palline da un'urna del genere, tutti i possibili risultati (e quindi lo spazio campione Ω) sono i quattro eventi seguenti:

$$\Omega = \{(B_1 \cap B_2),\ (B_1 \cap N_2),\ (N_1 \cap B_2),\ (N_1 \cap N_2)\}$$

Essi costituiscono una partizione (ovviamente non-equiprobabile) di Ω e – trattandosi di un'estrazione in blocco – possiedono le seguenti probabilità:

$$\begin{aligned}
Pr(B_1 \cap B_2) &= Pr(B_1)\,Pr(B_2|B_1) &&= \frac{3}{10}\frac{2}{9}\\
Pr(B_1 \cap N_2) &= Pr(B_1)\,Pr(N_2|B_1) &&= \frac{3}{10}\frac{7}{9}\\
Pr(N_1 \cap B_2) &= Pr(N_1)\,Pr(B_2|N_1) &&= \frac{7}{10}\frac{3}{9}\\
Pr(N_1 \cap N_2) &= Pr(N_1)\,Pr(N_2|N_1) &&= \frac{7}{10}\frac{6}{9}
\end{aligned}$$

Tenendo conto di questi risultati, è agevole determinare le probabilità degli eventi sopra elencati:

$$Pr(A) = Pr(N_1 \cap B_2) = \frac{7}{10}\frac{3}{9} = \frac{7}{30} = 0.23333;$$

$$\begin{aligned}
Pr(C) &= Pr((B_1 \cap N_2) \cup (N_1 \cap B_2)) = Pr(B_1 \cap N_2) + Pr(N_1 \cap B_2)\\
&= \frac{3}{10}\frac{7}{9} + \frac{7}{10}\frac{3}{9} = \frac{7}{15} = 0.46667;
\end{aligned}$$

$$\begin{aligned}
Pr(D) &= Pr((B_1 \cap B_2) \cup (N_1 \cap N_2)) = Pr(B_1 \cap B_2) + Pr(N_1 \cap N_2)\\
&= \frac{3}{10}\frac{2}{9} + \frac{7}{10}\frac{6}{9} = \frac{8}{15} = 0.53333;
\end{aligned}$$

$$\begin{aligned}
Pr(E) &= Pr((B_1 \cap N_2) \cup (N_1 \cap B_2) \cup (B_1 \cap B_2))\\
&= Pr(B_1 \cap N_2) + Pr(N_1 \cap N_2) + Pr(B_1 \cap B_2)\\
&= \frac{3}{10}\frac{7}{9} + \frac{7}{10}\frac{3}{9} + \frac{3}{10}\frac{2}{9} = \frac{8}{15} = 0.53333;
\end{aligned}$$

$$\begin{aligned}
Pr(F) &= Pr((N_1 \cap B_2) \cup (B_1 \cap B_2)) = Pr(N_1 \cap B_2) + Pr(B_1 \cap B_2)\\
&= \frac{7}{10}\frac{3}{9} + \frac{3}{10}\frac{2}{9} = \frac{3}{10} = 0.30000
\end{aligned}$$

L'ultimo risultato mostra che la probabilità dell'evento «la prima pallina estratta è Bianca» coincide con la probabilità dell'evento «la seconda pallina estratta è Bianca». Tale risultato è banale se l'estrazione è con ripetizione perché l'urna non si modifica ma, a prima vista, è inaspettato quando le sottoprove sono svolte senza ripetizione.

In effetti, bisogna accuratamente distinguere tra la probabilità $Pr(B_2|B_1)$ che la seconda pallina estratta è Bianca dato che la prima è Bianca e, invece, la probabilità dell'evento F ={la seconda pallina è Bianca, a prescindere dal colore della prima}. Infatti, *prima di svolgere le due estrazioni*, deve essere logicamente indifferente scommettere sul colore della prima pallina oppure sul colore della seconda. È l'informazione «si è verificato B_1» che modifica la probabilità di B_2 nel caso di estrazioni senza ripetizione.

Questa discussione ha una conseguenza importante per la teoria dei campioni: se si sceglie in modo casuale un gruppo di n persone da una popolazione di N, la probabilità che un individuo ha di essere prescelto (prima di iniziare le singole estrazioni) è costante sia che si operi con ripetizione che senza ripetizione. □

Esempio 8.15

Le *prove sequenziali* sono importanti nella verifica di un prodotto elettronico, di un'automobile, di un protocollo per la cura di una malattia, etc. A scopo didattico, i problemi verranno impostati in un modo molto semplice, cioè mediante estrazioni consecutive di palline da urne di cui si conosce la composizione. Modifiche opportune nelle modalità delle estrazioni generano situazioni più realistiche.

Supponiamo che un'urna U_1 contenga 3 palline Bianche e 7 Nere e che una seconda urna U_2 contenga 1 pallina Bianca e 9 Nere. La prova consiste in due sottoprove così articolate: si estrae una pallina dall'urna U_1 e se ne controlla il colore; se la pallina è Nera, la prova termina; se la pallina è Bianca, la si inserisce nell'urna U_2 da cui, dopo aver accuratamente rimescolato, si estrae una pallina.

Pertanto, alla fine, *la prova genera l'evento pallina Bianca* se e solo se esce pallina Bianca dalla prima urna e poi pallina Bianca dalla seconda urna; per contro, *la prova genera pallina Nera* se esce Nera dalla prima urna (e la prova termina), oppure se esce Bianca dalla prima urna e poi Nera dalla seconda urna.

Indichiamo con B l'evento «la prova genera pallina Bianca» e con N l'evento «la prova genera pallina Nera»; indichiamo, inoltre, con B_1, N_1 gli eventi generati da U_1 e con B_2, N_2 gli eventi generati da U_2. Allora, evidentemente,

$$B = (B_1 \cap B_2); \qquad N = (N_1) \cup (B_1 \cap N_2).$$

Le due sottoprove (cioè l'estrazione dalla prima urna e la successiva ed eventuale estrazione dalla seconda urna) non sono indipendenti perché il verificarsi di un evento nella prima estrazione modifica la probabilità dell'evento generato dalla seconda sottoprova. Quindi, si ottiene:

$$\begin{aligned}
Pr(B) &= Pr(B_1 \cap B_2) = Pr(B_1)\ Pr(B_2|B_1)\\
&= \frac{3}{10}\frac{2}{11} = \frac{3}{55} = 0.05455;
\end{aligned}$$

infatti, $Pr(B_2|B_1) = 2/11$ perché se si verifica Bianca nella prima estrazione tale pallina viene inserita nella seconda urna che, quindi, contiene, ora, 2 palline Bianche e 9 Nere. Similmente, si ha:

$$\begin{aligned}
Pr(N) &= Pr((N_1) \cup (B_1 \cap N_2)) = Pr(N_1) + Pr(B_1)\,Pr(N_2|B_1)\\
&= \frac{7}{10} + \frac{3}{10}\frac{9}{11} = \frac{52}{55} = 0.94545
\end{aligned}$$

Ovviamente, poteva anche dedursi direttamente: $Pr(N) = 1 - Pr(B)$ perché gli eventi B e N sono necessari, per cui la somma delle loro probabilità deve essere uguale ad 1.

Lo schema proposto si può applicare e generalizzare in molti modi. Per esempio, in una prova di esame che consiste in un compito scritto ed in un colloquio orale, il candidato viene ammesso all'orale se e solo se supera la prova scritta (pallina Bianca). In prove sequenziali di resistenza a dosi crescenti di farmaci, le cavie vengono sottoposte agli esperimenti successivi solo se superano quelli precedenti. Nei tornei ad eliminazione diretta, solo le squadre vincitrici accedono alle partite successive, e così via. □

Esempio 8.16

Un aspetto connesso all'*attesa di un evento* concerne il problema dei «ritardi nel gioco del Lotto», i quali generano spesso credulità e aspettative ingiustificate. La questione paradossale deriva dal fatto che, mentre ciascuno accetta l'indipendenza delle successive estrazioni, pochi concludono coerentemente che la probabilità di un evento qualsiasi in una certa estrazione «non può essere modificata» da qualsiasi evento verificatosi in una precedente estrazione. In effetti, dovrebbe essere abbastanza ovvio che le sottoprove successive «non hanno memoria» degli eventi passati.

La credulità è molto diffusa, anche in settori «colti» della popolazione. Eppure, basterebbe chiedersi perché gli studiosi di Probabilità e di Statistica non abbandonino il loro mestiere per investire scientificamente in tali studi. Più prosaicamente, perché decine di riviste derivano il loro reddito dalle copie vendute a lettori molto creduloni (che cercano in esse combinazioni sicure basate su tali ritardi) e non si arricchiscono più semplicemente utilizzando in proprio tali conoscenze? Purtroppo, tale errata opinione viene rafforzata dalle informazioni parziali diffuse dai mass-media: per esempio, si pubblicizza con enfasi che un ritardo lungamente atteso ha prodotto milioni di Euro di vincite complessive, dimenticando di aggiungere, però, che durante l'attesa si è versato complessivamente allo Stato un contributo di gran lunga superiore alle vincite occasionalmente elargite.

In modo formale, se θ è la probabilità che un prefissato numero si verifichi ad una data estrazione del Lotto, si tratta di calcolare la probabilità che, non essendosi verificato per $m-1$ estrazioni consecutive, il numero si verifichi alla m-esima.

Ora, se indichiamo con E_j l'evento che il numero si verifichi all'estrazione j-esima, la probabilità richiesta è espressa dalla probabilità condizionata:

$$\begin{aligned} Pr\left(E_m|\overline{E}_1 \cap \overline{E}_2 \cap \cdots \cap \overline{E}_{m-1}\right) &= \frac{Pr\left(\overline{E}_1 \cap \overline{E}_2 \cap \cdots \cap \overline{E}_{m-1} \cap E_m\right)}{Pr\left(\overline{E}_1 \cap \overline{E}_2 \cap \cdots \cap \overline{E}_{m-1}\right)} \\ &= \frac{(1-\theta)^{m-1}\theta}{(1-\theta)^{m-1}} = \theta = Pr\left(E_m\right). \end{aligned}$$

Quindi, la conoscenza del «lungo ritardo» non ha modificato – come era necessario – la probabilità (costante) che un prefissato numero si verifichi alla successiva estrazione.

Probabilmente, l'illusione che abbia un senso razionale la scommessa sui numeri «ritardari» in un gioco a sottoprove indipendenti, deriva dalla confusione tra due eventi che, nel linguaggio comune, si reputano coincidenti ma che, invece, sono completamente differenti.

Infatti, spesso, si sente dire che è praticamente impossibile che un numero non esca su una medesima «ruota del Lotto» per (diciamo) 400 estrazioni, «quindi» deve verificarsi alla 401-esima estrazione con altissima probabilità.

Ora, è vero che è quasi 0 la probabilità che un numero non venga estratto in una lunghissima sequenza di sottoprove, *ma il giocatore del Lotto non scommette su questo evento, bensì sull'evento connesso alla singola estrazione*, e la cui probabilità resta immutata, perché trattasi di sottoprove indipendenti, come abbiamo appena evidenziato. □

Esempio 8.17

L'esempio che ora discuteremo genera una certa sorpresa nel lettore perché, quasi sempre, ci si attenderebbe una probabilità più bassa di quella che si ottiene con il calcolo; ciò deriva dal fatto che il nostro cervello ha una notevole difficoltà nel percepire l'ordine di grandezza delle combinazioni tra eventi elementari. Perciò, spesso, perveniamo ad una valutazione per difetto della probabilità di eventi complessi. Inoltre, vi è la tendenza a considerare eccezionali eventi che invece possono avere probabilità non-trascurabili.

Supponiamo che in una certa stanza vi siano n persone e ci si chiede qual è la probabilità che almeno 2 festeggino il *compleanno* lo stesso giorno (si noti che non si richiede di avere la stessa età, ma solo lo stesso giorno e lo stesso mese di nascita).

Ora, poiché ciascuno può esser nato in uno dei 365 giorni dell'anno – escludiamo gli anni bisestili – lo spazio campione Ω è composto da tutti i possibili giorni di nascita delle n persone, e quindi consta di $(365)(365)\ldots(365) = 365^n$ eventi elementari.

Se ipotizziamo che ciascuno possa essere nato in un giorno qualsiasi con la stessa probabilità, allora ci troviamo di fronte ad una *partizione equiprobabile* di Ω. Infatti, i 365^n eventi elementari sono:

i) *necessari*, perché sicuramente le n persone saranno nate in una delle sequenze di Ω;

ii) *incompatibili*, perché nessuno può essere nato in due date differenti;

iii) *equiprobabili*, il che non è generalmente vero, ma questa ipotesi semplificatrice consente di rispondere agevolmente alla domanda.

Quindi, è corretto assegnare ad ogni evento E come probabilità il rapporto ottenuto calcolando il numero delle sequenze di eventi elementari di cui si compone E ed il numero totale delle sequenze elementari di cui si compone Ω.

Per calcolare la probabilità dell'evento:

E_n={almeno una coppia tra n persone festeggia il compleanno lo stesso giorno}

è più semplice enumerare gli eventi elementari favorevoli all'evento complementare:

$\overline{E}_n$={nessuna coppia tra n persone festeggia il compleanno lo stesso giorno}.

Se nella stanza vi sono solo $n = 2$ persone, il numero di eventi elementari affinché $\overline{E}_n$ si verifichi è $(365)(364)$, perché essi sono tutti i possibili 365 giorni di nascita della prima persona moltiplicati per i rimanenti possibili 364 giorni di nascita della seconda persona *se è nata in un giorno differente dalla prima.*

Se le persone sono $n = 3$, il numero di eventi affinché $\overline{E}_n$ si verifichi sono $(365)(364)(363)$ perché queste sono tutte e sole le possibili triple di giorni di nascita tali che *ciascuna persona sia nata in un giorno diverso dalle altre due.*

Generalizzando per ogni intero $n \in [1, 365]$, il numero di eventi elementari favorevoli a $\overline{E}_n$ è pari a: $(365)(364)...(365-n+1)$.

Pertanto, per ogni $n = 1, 2, \ldots, 365$, si ha:

$$Pr\left(E_n\right) = 1 - Pr\left(\overline{E}_n\right) = 1 - \frac{(365)(364)...(365-n+1)}{365^n};$$

Ovviamente, $Pr(E_n) = 1$, per $n \geq 366$; infatti, se vi sono più di 365 persone è matematicamente certo che almeno 1 persona è nata nello stesso giorno di un'altra.

Molti credono che occorra un numero n piuttosto elevato di persone perché la probabilità $Pr(E_n)$ risulti elevata. Invece, calcolando tale probabilità per alcuni valori di n (come mostrato in dettaglio nel paragrafo R 8.2), si hanno i valori seguenti:

$n =$	10	20	23	30	40	50	60	70
$Pr(E_n) =$	0.117	0.411	0.507	0.706	0.891	0.970	0.994	0.999

Come si vede, tale probabilità è già elevata per poche decine di persone; per esempio, $Pr(E_n) > 0.5$, se $n \geq 23$. Poiché tale conclusione è inattesa, se l'evento si verifica lo si definirebbe una *coincidenza notevole*. Invece, se in un gruppo di 23 persone una coppia scopre di festeggiare il compleanno lo stesso giorno si è verificata una «coincidenza» (!) quasi uguale a quella del verificarsi di una «Testa» lanciando una moneta ben equilibrata.

Nel problema dei compleanni si è ipotizzato che sia costante la probabilità di nascere in ciascuno dei 365 giorni dell'anno, il che non è realistico sulla base degli studi demografici. Ebbene, si può dimostrare che se non vi è equiprobabilità delle nascite nei 365 giorni dell'anno (come in effetti avviene) la probabilità di una «coincidenza» aumenta. Ciò è ragionevole perché se, per assurdo, tutti nascessero lo stesso giorno, l'evento E_n avrebbe probabilità 1.

□

8.15 Considerazioni finali

In questo paragrafo finale, ci interessa svolgere alcune considerazioni che emergono dallo studio della probabilità e dalla sua centralità rispetto ad un percorso unitario di analisi statistica della realtà. Esse riguarderanno, nell'ordine, l'esigenza di unificare le varie interpretazioni della probabilità, i legami tra il Calcolo delle probabilità e la Statistica esplorativa (presentata nella prima parte) e le conseguenze delle nozioni probabilistiche sin qui svolte per l'Inferenza statistica (che affronteremo nella terza parte).

■ Il fatto che ciascuna proposta, avanzata per l'interpretazione delle origini concettuali della probabilità e per la sua conseguente misura, possieda una validità teorica ed una valenza applicativa suggerisce un approccio ecumenico nel quale convergano e convivano tutti i risultati giusti e coerenti delle varie interpretazioni. In tal modo, ogni persona – nello specifico delle situazioni e delle esigenze connesse alla sua prova particolare – potrà decidere in modo ottimale.

Anzitutto, *non si può non essere soggettivisti* essendo ben evidente che ciascuno di fronte agli stessi fatti offre interpretazioni differenti che gli derivano da un contenuto informativo diverso. I dibattiti politici su situazioni oggettive, le opinioni spesso divergenti sulle ragioni di un comportamento specifico, la forza dell'evidenza di prove processuali, e così via, sono ambiti di dibattito e di scontro nei quali si assiste quotidianamente a quantificazioni differenti (spesso inconscie) delle probabilità degli stessi esiti di una prova. D'altra parte, ogni soggetto razionale matura le sue convinzioni e giunge a decisioni conseguenti sulla base del ragionamento e/o dell'esperienza. Quindi, in modo implicito o esplicito, a parere nostro, ogni valutazione soggettivista della probabilità discende – a seconda dei casi – dalla definizione classica (frutto del ragionamento) e/o da quella frequentista (frutto della esperienza); e ciò non solo nella formulazione personale dei convincimenti ma anche per sostenere con forza le proprie convinzioni in modo oggettivo e dimostrabile anche ad altri, formulando argomentazioni che derivano dalla similitudine, dalla analogia o dal contrasto con esperimenti già realizzati o solo ipotetici. È indubbio che nella vita quotidiana come nella ricerca delle cause o nel confronto di tesi o soluzioni alternative vi sia un atteggiamento psicologico assai simile a quello di chi scommette: il fatto è che, nelle scommesse, la valutazione numerica sembra essere un dato derivato dal ragionamento e/o dall'esperienza e non un dato primario. E questo anche a prescindere dalla ineliminabile valutazione utilitaristica sempre presente in ogni decisione umana.

Per questo, ci pare ugualmente corretto affermare che *non si può non essere frequentisti* nella misura in cui la realtà – nella replica dei fatti e nella loro variabilità – ci suggerisce leggi, teorie, comportamenti e valutazioni. Il *postulato empirico del caso*, secondo cui «le frequenze relative di un evento convergono alle rispettive probabilità», e le leggi dei grandi numeri che, in vari contesti, apportano un sostegno matematico a tale intuizione, derivano da constatazioni sperimentali impresse così fortemente e con naturalezza nel nostro ragionamento che tutti i nostri comportamenti e le nostre valutazioni sono, spesso anche se non esclusivamente, dettate da una logica frequentista. Naturalmente, esistono esperimenti irripetibili di fronte ai quali la valutazione probabilistica non può essere condotta in modo oggettivo; anche in tali casi, pur aderendo ad una interpretazione che vede la determinazione della probabilità come il risultato di una scommessa, crediamo che l'Uomo pervenga ad un proprio convincimento probabilistico secondo logica, raziocinio, deduzione, similitudine, contrasto e analogia: i quali sono tutti momenti del pensare umano che richiamano altri esperimenti o da essi derivano.

La intrinseca unitarietà e l'utilità operativa di una convivenza pragmatica tra le differenti impostazioni della probabilità ci viene offerta dalla semplice constatazione che l'evoluzione umana e la sopravvivenza quotidiana si sono realizzate, e tuttora si realizzano, beneficiando di una sostanziale regolarità stocastica degli eventi «ripetibili», «quasi ripetibili» e «grossolanamente ripetibili». In tale accezione, il postulato empirico del caso non è solo una ipotesi ma una necessità ontologica. Infatti, a parere nostro, se *per assurdo* i risultati di una prova si presentassero secondo successioni e modalità tali da contraddire la stabilità degli esiti noi lo potremmo rilevare se e solo se questa instabilità fosse regolare al punto da essere accertata con qualche grado di sicurezza, cioè solo se anche per essa valesse un postulato empirico del caso.

In breve, siamo fermamente convinti che la valutazione probabilistica di un evento è una operazione complessa che avviene nel cervello umano sulla base di una molteplicità di informazioni che derivano da fattori oggettivi connessi alla prova, da esperienze precedenti, da valutazioni personali. Esistono prove nelle quali è più netta la prevalenza verso il primo atteggiamento e quindi generano una valutazione unanime; ne esistono altre per le quali le esperienze non sono così concordanti per cui possono esistere discrepanze non sostanziali nelle differenti valutazioni; altre infine assurgono al rango di situazioni uniche di fronte alle quali l'opinione personale resta l'unica possibilità di fornire valutazioni probabilistiche. Ciò che conta, però, non è l'assolutezza delle posizioni ma solo una convivenza tra approcci che, per quanto è dato di riscontrare nella vita di ogni giorno, produce sempre soluzioni convincenti e atteggiamenti pragmatici di grande utilità.

■ Per collocare il Calcolo delle probabilità all'interno di quel percorso unitario che caratterizza questo volume, occorre sottolineare il legame tra questo

Capitolo e le precedenti discussioni sulla Statistica esplorativa, anticipando poi alcune questioni che si ritroveranno nella parte terza.

Supponiamo che si intervisti una classe di 20 studenti per valutare l'atteggiamento verso una proposta e che la distribuzione delle frequenze assolute risulti essere la seguente: 8 favorevoli, 3 contrari, 9 indifferenti. Quindi, la variabile statistica (qualitativa) X =«atteggiamento verso il prodotto», in rapporto ai tre possibili attributi, presenta la seguente distribuzione delle frequenze relative:

Attributi	Frequenze relative
favorevole	$8/20 = 0.40$
contrario	$3/20 = 0.15$
indifferente	$9/20 = 0.45$

Ora, nella medesima classe, definiamo una prova $\mathcal{P}$ scegliendo a caso uno studente: per esempio, inserendo i 20 nomi su dei foglietti ripiegati in una scatola, mescolando accuratamente, ed estraendo poi un foglietto e chiedendo allo studente il cui nome è stato così prescelto il suo atteggiamento verso il prodotto. Si chiede, allora, di calcolare qual è la probabilità che si verifichi uno dei tre possibili eventi: «lo studente estratto è favorevole», «lo studente estratto è contrario», «lo studente estratto è indifferente». Poiché la prova genera 1 risultato sui 20 possibili, tutti necessari incompatibili ed equiprobabili, è possibile applicare la regola pratica dedotta nel paragrafo 8.12 ed ottenere le seguenti assegnazioni di probabilità:

Eventi	Probabilità
favorevole	$8/20 = 0.40$
contrario	$3/20 = 0.15$
indifferente	$9/20 = 0.45$

Cosa differenzia, allora, il Calcolo delle probabilità dalla Statistica descrittiva? È *la prova, cioè l'esperimento soggetto ad incertezza, che trasforma una distribuzione di frequenze relative in una assegnazione coerente di probabilità su eventi discreti.*

Quindi, tutte le volte che su una distribuzione di frequenze relative osservate si istituisce una prova si definisce anche uno spazio di probabilità discreto ed una assegnazione corretta di probabilità; viceversa, tutte le volte che si opera una assegnazione di probabilità ad una collezione di eventi questa può sempre essere interpretata come la distribuzione di frequenze relative delle corrispondenti modalità.

In sintesi, la Statistica descrittiva esamina i risultati di esperimenti reali, già avvenuti e definitivi, di cui studia *a posteriori* la distribuzione del carattere X tra le singole modalità; invece, il Calcolo delle probabilità elenca i risultati di esperimenti ipotetici, che non necessariamente si realizzano, di cui esamina *a priori* le differenti probabilità dei singoli eventi.

La similitudine tra i due aspetti sarà molto evidente nel Capitolo 9, allorquando per risolvere il problema della sintesi dei risultati di una prova si utilizzeranno praticamente tutti gli indicatori sintetici di posizione, variabilità e forma, prima introdotti per le distribuzioni di frequenza.

■ Infine, vogliamo rimarcare come il concetto e la misura della probabilità giochino un ruolo essenziale per lo sviluppo rigoroso dell'Inferenza statistica cui saranno dedicate le parti terza e quarta di questo volume. Perché sia evidente e semplice, lo faremo sviluppando un esempio concreto dal quale si possono intuire alcune problematiche inferenziali.

Supponiamo di incontrare un vecchio amico che ci invita a giocare «a Testa e Croce» estraendo dalla sua tasca una moneta e invitandoci a puntare sull'evento «Croce». Se dopo 5 lanci esce sempre Testa rimarremo un po' perplessi perché siamo abituati ad attenderci un certo equilibrio tra il numero di eventi «Testa» e «Croce».

Tuttavia, dopo 5 lanci, non metteremo in discussione la lealtà del nostro amico perché la probabilità che si verifichino consecutivamente 5 «Testa» in una moneta ben equilibrata è $1/32 = 0.03125$, e tale probabilità – per quanto bassa – ci sembra troppo alta per incrinare una vecchia amicizia e offendere un amico.

Tale sicurezza, però, comincia a vacillare se per 10, 20, 30 lanci si verifica sempre «Testa»; infatti, la probabilità di tali eventi è, rispettivamente,

$$\frac{1}{1024} = 0.97656 \times 10^{-3}; \frac{1}{1.048.576} = 0.95367 \times 10^{-6}; \frac{1}{1.073.741.824} = 0.93132 \times 10^{-9},$$

ed è così bassa da consentirci legittimamente di dubitare della correttezza dei lanci della moneta (o, addirittura, che essa sia costruita con due facce «Testa»!).

Naturalmente, possono *effettivamente* presentarsi 10, 20 e anche 30 volte «Testa» di seguito, con una moneta perfettamente equilibrata, ma il nostro ragionamento ci condurrà con maggiore naturalezza a dubitare fortemente piuttosto che ammettere di essere testimone di una sequenza di risultati molto rara.

Questo esempio evidenzia un problema generale della ricerca scientifica: esiste una ipotesi H_0 (il mio amico è onesto, la moneta è ben equilibrata) che io ritengo valida fino a quando non si verifica un evento molto raro (che cioè presenta probabilità inferiore a $10^{-3}, 10^{-6}, 10^{-9}$).

Tra l'accogliere l'eventualità che si sia verificato un evento molto raro, ma possibile, e il mettere in discussione l'ipotesi H_0 rifiutandola a favore di qualche ipotesi alternativa (la moneta è pesantemente truccata a favore dell'evento «Testa», l'amico utilizza un sistema di lancio che privilegia sistematicamente la faccia «Testa», la moneta contiene due facce «Testa») si preferisce mantenere valida H_0 sino a che, sussistendo H_0, l'evento non ha una probabilità talmente bassa da ritenerla ingiustificata.

Va sottolineato, come l'esempio conferma, che qualsiasi decisione si assuma con riguardo ad H_0 (la moneta è corretta) essa ha dei rischi, e quindi dei costi: perdere soldi se continuo a credere in H_0 quando invece H_0 è falsa; perdere l'amicizia se rifiuto H_0 quando invece H_0 è vera.

Poiché, per definizione, *ogni decisione statistica è soggetta ad incertezza*, emerge il ruolo determinante della probabilità, assieme al costo-utilità delle decisioni da assumere, per individuare livelli accettabili di rischio.

Questo aspetto sarà centrale per lo sviluppo delle differenti problematiche dell'Inferenza statistica e, come si vede, riposa tutto sulla Teoria delle probabilità.

Concludiamo questo Capitolo con una convinzione che appare sempre più diffusa: «la probabilità è una delle più grandi conquiste del XX secolo».

Pur avendo regolarmente accompagnato l'avventura umana, nella vita quotidiana come nella ricerca di laboratorio, questo concetto primitivo ha manifestato il suo ruolo centrale ed essenziale per le decisioni correnti solo nella fase matura dello sviluppo intellettuale dell'Umanità.

Dopo la perdita di quelle numerose certezze, che hanno contribuito in modo determinante allo sviluppo della nostra specie, la Scienza si evolve oggi grazie

ad una conoscenza critica, nel pensiero e nella ricerca, che è necessariamente di tipo statistico: e questa diffusa consapevolezza deriva la sua ragion d'essere e la sua efficacia operativa da una comprensione della realtà che si esprime in termini probabilistici.

Ambiente R

Calcolo delle probabilità

R8.1 Estrazione di palline da un'urna

Le elaborazioni necessarie per il Capitolo 8 si riducono ad operazioni elementari sulle frazioni, per cui non richiedono un particolare utilizzo di R (se non come semplice calcolatore numerico). Per questo, ci limitiamo ad esemplificare poche situazioni, iniziando dall'estrazione di palline colorate da una scatola, in analogia ai numerosi esercizi di questo tipo presenti nel testo.

L'estrazione di oggetti da un'urna (con o senza ripetizione) avviene in R con la funzione sample(), con le opzioni esemplificate nelle seguenti istruzioni commentate:

```
### estrae SENZA ripetizione per 10 volte numeri da 1 a 10
R==> sample(1:10);
[1]  8  7  6  2  1  3  9  4  5 10

### estrae CON ripetizione  per 10 volte i numeri da 1 a 10
R==> sample(1:10,replace=TRUE);
[1]  4  4  6  6  9 10  9  8  8  7

### lancia una moneta ben equilibrata per 100 volte
R==> sample(c("Testa", "Croce"), 100, prob=c(0.5,0.5),
             replace=TRUE);
  [1] "Croce" "Croce" "Testa" "Croce" "Croce" "Croce" "Testa"
             "Testa" "Testa"
 [10] "Croce" "Croce" "Testa" "Croce" "Testa" "Croce" "Croce"
             "Testa" "Croce"
 [19] ....................................................
[100] "Testa"

### estrai per 30 volte CON ripetizione una pallina da un'urna
### che contiene palline: (3 Bianche, 5 Rosse, 2 Verdi)
R==> sample(c("B","R","V"), 30, prob=c(3/10,5/10,2/10),
             replace=TRUE);
 [1] "R" "V" "R" "R" "B" "V" "R" "R" "V" "R" "R" "V" "V" "R" "R"
[16] "R" "V" "R" "V" "B" "R" "B" "B" "R" "R" "R" "B" "V" "R" "R"

### estrai il numero d'ordine di 10 studenti SENZA RIPETIZIONE
```

```
### dal data set studenti per una scelta casuale delle unità.
### La numerosità della popolazione viene calcolata
### con length(genere).
R==> dati=read.table("studenti.txt", header=TRUE);
R==> with(data=dati, sample(1:length(genere),10));
[1]  924  354  361 1419  955 1094 2200  535  105   84
```

Si osservi che se nulla è detto la funzione `sample()` opera *senza ripetizione*; quindi, estraendo n elementi da un gruppo di n si produce una permutazione casuale degli n elementi (perché vengono estratti tutti in ordine casuale). Tale funzione sarà ri-utilizzata parlando di simulazione nei Capitoli 11 e 12.

R8.2 Problema del compleanno

Nell'Esempio 8.17 del testo è proposto e risolto il cosiddetto "problema dei compleanni" (=*birthday's problem*) nel quale la probabilità di un evento è più alta di quella che, comunemente, la gente crede.

Per evidenziare come tale probabilità cresca rapidamente con il numero dei presenti (indicati con n), calcoliamo e rappresentiamo graficamente la funzione:

$$Pr(E_n) = 1 - \frac{365 \times 364 \times \cdots \times (365-n+1)}{365^n} = 1 - \prod_{k=1}^{n} \frac{365-k+1}{365} = 1 - \prod_{k=1}^{n} \left(1 - \frac{k-1}{365}\right)$$

al variare di $n = 1, 2, \ldots, 366$. L'ultima formulazione è più efficiente dal punto di vista computazionale perché si opera su frazioni proprie; invece, la prima espressione implica il rapporto di numeri che diventano molto elevati al crescere di n.

Tale probabilità varia tra 0 (quando $n = 1$) ed 1 (quando $n = 366$). Per ottenerla per ciascun valore di n, si utilizza la funzione `cumprod()` che fornisce il prodotto cumulato, cioè il prodotto dei termini consecutivi al variare di n. Quindi, per calcolare tali probabilità le istruzioni sono le seguenti:

```
R==> vett=1:366;
R==> probn=1-cumprod(1-(0:365)/365);
```

Per mostrare come tale probabilità varia con n, costruiamo un grafico (Figura R 8.1) per ogni n. Le istruzioni seguenti:

```
R==> par(mfrow=c(2,1))      ### suddividi l'area grafica
                            ### in 2 regioni
R==> plot(vett,probn,xlab="n= Numero dei presenti",
       ylab="Probabilità");
R==> plot(vett[1:70],probn[1:70],xlab="n= Numero dei presenti"
       ,ylab="Probabilità");
R==> par(mfrow=c(1,1))      ### ripristina l'area grafica
```

generano il grafico di sinistra che mostra la probabilità per $n = 1, 2, \ldots, 366$ mentre quello di destra ne mostra il dettaglio per $n = 70$; in effetti, per $n > 70$, tale probabilità è superiore a 0.999.

R8.3 Un problema di Calcolo delle probabilità

Da una scatola che contiene palline numerate da 1 a n se ne estraggono 3 senza ripetizione e si chiede qual è la probabilità che esse siano crescenti, cioè che la prima

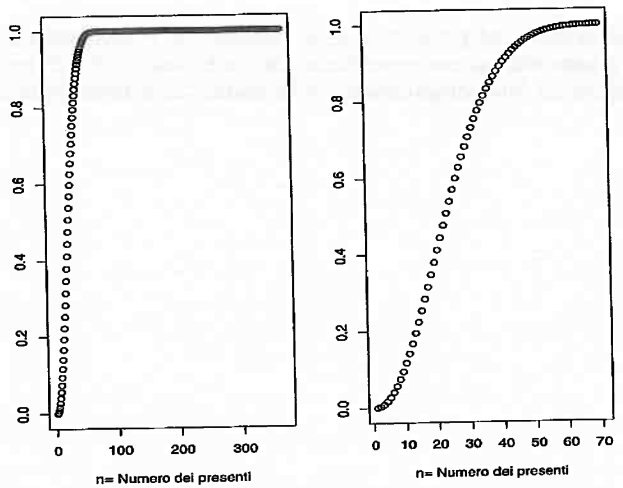

FIGURA R8.1. *Probabilità che tra n persone almeno due festeggino il compleanno lo stesso giorno.*

estratta sia inferiore alla seconda che sia inferiore alla terza. Quindi, l'evento E definito come "i tre numeri estratti sono crescenti" si verifica se sono estratte, per esempio, le terne $(1, 2, 8)$ oppure $(2, 3, 10)$, ... ma non si verifica per $(2, 1, 3)$.

Poiché la prova genera una partizione equiprobabile di eventi elementari, occorre contare tra tutti gli eventi elementari possibili quelli favorevoli ad E (nel senso che soddisfano il requisito richiesto): in tal caso, la probabilità si misura mediante il rapporto fra la numerosità degli eventi elementari favorevoli e quelli totali.

Di tale problema presentiamo sia una soluzione numerica che formale. In ogni caso, lo spazio campione consiste di $n(n-1)(n-2)$ eventi elementari mentre una generica terna (x_i, x_j, x_k) di numeri interi compresi tra 1 e n soddisfa il requisito $(x_i < x_j < x_k)$ di essere ordinatamente crescenti se e solo se si verifica, *contemporaneamente*, sia $(x_i - x_j) < 0$ che $(x_j - x_k) < 0$. A tal fine, si implementano i seguenti cicli annidati (perché ciascuno è inserito nel precedente) il cui scopo è, appunto, quello di "contare" quante sono le terne per le quali tale condizione si verifica. Tale algoritmo, riportato per $n = 1, 2, \ldots, 15$, non è particolarmente efficiente (infatti, per n elevato richiederebbe tempi molto lunghi) ma viene qui proposto per finalità didattiche.

```
R==> maxn=15;
R==> risult=matrix(NA,nrow=maxn,ncol=4);
R==> for(n in 3:maxn){
          totali=n*(n-1)*(n-2);
          quanti=0;
          for(i in 1:n){
          for(j in 1:n){
          for(k in 1:n){
              quanti=quanti+((i-j<0) & (j-k<0));
                        }
                        }
                        }
```

```
            prob=quanti/totali;
            risult[n,]=t(c(n,quanti,totali,prob));
            }
R==> risult;
      [,1] [,2] [,3]    [,4]
 [1,]   NA   NA   NA      NA
 [2,]   NA   NA   NA      NA
 [3,]    3    1    6 0.16667
 [4,]    4    4   24 0.16667
 [5,]    5   10   60 0.16667
 [6,]    6   20  120 0.16667
 [7,]    7   35  210 0.16667
 [8,]    8   56  336 0.16667
 [9,]    9   84  504 0.16667
[10,]   10  120  720 0.16667
[11,]   11  165  990 0.16667
[12,]   12  220 1320 0.16667
[13,]   13  286 1716 0.16667
[14,]   14  364 2184 0.16667
[15,]   15  455 2730 0.16667
```

Si osservi come, per ottenere una matrice di dimensioni $(maxn, 4)$ nella quale collocare, per ciascuna riga a partire dalla terza, i quattro valori di interesse, sia stato opportuno definire – prima dei vari *loop* – una matrice `risult` identicamente pari a NA. Presentando, alla fine, i risultati questa scelta aiuta a controllare la corretta collocazione dei risultati.

Sul piano numerico, è abbastanza evidente che la probabilità dell'evento E è costante ed è: $P_r(E) = \dfrac{1}{6} \simeq 0.16667$, per ogni $n \geq 3$.

Sul piano formale, per ottenere il medesimo risultato (senza utilizzare argomenti di analisi combinatoria non sviluppati nel testo) proponiamo un ragionamento elementare che suggeriamo al lettore di ripetere con foglio e penna. In breve, occorre contare le sole terne che soddisfano la condizione "numeri crescenti", cioè il numero di eventi elementari favorevoli ad E tra le $n(n-1)(n-2)$ partizioni equiprobabili generate dalla prova in questione. Evidentemente, le terne favorevoli possono avere come primo estratto soltanto i numeri $\{1, 2, \ldots, n-2\}$.

Ebbene, tra le terne favorevoli consideriamo quelle che cominciano con 1; il secondo estratto può essere solo un numero in $\{2, 3, \ldots, n-1\}$. Consideriamo, per un momento quelle la cui coppia iniziale sia $(1,2)$: ve ne sono $n-2$ favorevoli, e cioè: $(1,2,3)$, $(1,2,4), (1,2,5), \ldots, (1,2,n)$. Se consideriamo, poi, quelle la cui coppia iniziale è $(1,3)$ ve ne saranno $n-3$ favorevoli, e precisamente: $(1,3,4)$, $(1,3,5)$, $\ldots$, $(1,3,n)$. Continuando in questo modo (avendo per ora fissato come 1 il primo elemento della terna) si arriva all'ultima coppia $(1, n-1)$ per la quale affinché sia favorevole è possibile 1 solo caso e cioè la terna $(1, n-1, n)$. Alla fine, riassumendo tutte le terne favorevoli che cominciano con 1 esse sono la somma di $n-2$ (che contengono 2 al secondo posto) più $n-3$ (che contengono 3 al secondo posto), $\ldots$, più 1 (che contiene $n-1$ al secondo posto), per cui la numerosità di tali terne è pari a $S_1 = (n-2) + (n-3) + \cdots + 1$.

Successivamente, si tratta di considerare le terne favorevoli che cominciano con 2 per le quali il secondo estratto può essere solo un numero in $\{3, 4, \ldots, n-1\}$. Ripetendo il precedente ragionamento, è agevole concludere che tutte le terne favorevoli che cominciano con 2 sono la somma di $n-3$ terne (che contengono 3 al secondo posto) più $n-4$ terne (che contengono 4 al secondo posto), $\ldots$, più 1 terna (che contiene $n-1$ al secondo posto), per cui la numerosità di tali terne è: $S_2 = (n-3) + (n-4) + \cdots + 1$.

Proseguendo in tale modo, e facendo variare sempre il primo estratto, si arriva alle terne che cominciano con $n-3$ per le quali il secondo estratto può essere solo un numero in

$\{n-2, n-1\}$. Ed è agevole convincersi che vi sono solo 2 terne favorevoli che cominciano con $(n-3, n-2)$ ed 1 terna favorevole che comincia con $(n-3, n-1)$. Formalizzando come in precedenza, la numerosità di tali terne è: $S_{n-3} = 2 + 1$.

Infine, l'unica terna favorevole che inizia con $n-2$ è $(n-2, n-1, n)$ e, quindi, $S_{n-2} = 1$.

Allora, per conoscere quante sono le terne favorevoli all'evento E occorre sommare tali numerosità:

$$
\begin{aligned}
& S_{n-2} + S_{n-3} + \cdots + S_2 + S_1 \\
= \; & [1] + [1+2] + \cdots + [1+2+\cdots+(n-4)+(n-3)] + [1+2+\cdots+(n-3)+(n-2)] \\
= \; & \sum_{j=1}^{n-2} j\,(n-j-1) = (n-1)\sum_{j=1}^{n-2} j - \sum_{j=1}^{n-2} j^2 \\
= \; & (n-1)\,\frac{(n-2)(n-1)}{2} - \frac{(n-2)(n-1)(2n-3)}{6} = \frac{n(n-1)(n-2)}{6}
\end{aligned}
$$

nei cui sviluppi abbiamo utilizzato le formule che sommano numeri interi ed i lori quadrati, cioè:

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}\,; \qquad \sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}\,.$$

Alla fine, la probabilità dell'evento E richiesto è:

$$P_r(E) = \frac{n(n-1)(n-2)/6}{n(n-1)(n-2)} = \frac{1}{6}\,.$$

come avevamo già trovato sul piano numerico.

Concludiamo, osservando che la maggior parte delle persone interpellate –prima di svolgere questo problema– non pensa che il risultato sia costante al variare di n.

Capitolo nono

Variabili casuali

9.1 Introduzione

L'assegnazione di una misura della probabilità per gli eventi (intesi come risultati delle prove, secondo le linee sviluppate nel capitolo precedente) non è sufficiente per risolvere la gran parte dei problemi reali che si presentano con connotati di incertezza. Ciò avviene quando si è interessati ad una sintesi degli eventi oppure quando occorre pervenire ad un ordinamento tra strategie e scelte, ciascuna delle quali ad esito incerto.

- Se si intervistano 30 persone da una popolazione molto ampia per conoscere se sono favorevoli ad una certa legge, non si è certamente interessati alle $2^{30} = 1.073.741.824$ possibili sequenze di «favorevole, contrario» cui tali prove possono dare luogo. In effetti, nessuno è particolarmente interessato a fornire una probabilità per tutti i possibili risultati di una tale prova mentre interessa molto di più la proporzione dei favorevoli alla legge in questione. Un problema simile si verifica dopo l'esito di una cura con le modalità «guarito, non-guarito»; dopo l'esito di una scommessa con le modalità «vinto, perso», etc.
- In altre situazioni occorre scegliere tra alternative soggette ad incertezza e rispetto alle quali si è interessati ad una valutazione sintetica dei risultati. Per esempio, dovendo scegliere tra itinerari stradali alternativi, differenti titoli quotati in Borsa, una pluralità di percorsi di studio o più modalità di prove di esame, etc. si desidera ottimizzare – rispettivamente – il tempo di percorrenza, i guadagni settimanali, la votazione finale, etc. In tali casi, non essendo certi degli esiti delle differenti prove, occorre graduare la scelta mediante il concetto di «aspettativa» secondo cui, dovendo prendere una decisione tra più scelte comparabili, si preferisce quella che, nel lungo periodo, fornisce un risultato complessivamente più favorevole al decisore.

È, quindi, necessario introdurre sullo spazio degli eventi una relazione algebrica e una relazione d'ordine, il che avviene mediante la definizione di *variabile casuale* (abbreviata con v.c., nel seguito). A tal fine, nel seguito, introdurremo le due tipologie di v.c. (discrete e continue) proponendo poi alcune misure di sintesi le quali, per molti aspetti, richiamano quelle già introdotte per le distribuzioni di frequenza nei precedenti capitoli.

Preliminarmente, rispetto ad ogni v.c., occorre porsi le seguenti domande:

- *quando possiamo dire di conoscere una v.c.?*
- *sotto quali condizioni una v.c. è ben definita?*
- *come possiamo rappresentare graficamente una v.c.?*

Nei prossimi paragrafi affronteremo la problematica delle v.c. distinguendo subito tra v.c. discrete e continue. Tale distinzione è di fondamentale importanza per il prosieguo della trattazione.

9.2 Definizione di variabile casuale

In termini generali ed intuitivi, una variabile casuale (univariata) è una regola che associa ad ogni evento un unico numero reale.

Formalmente, *una **variabile casuale** è una funzione misurabile e a valori reali definita sullo spazio* Ω.

o Il problema di definire esattamente cosa si intende per *funzione misurabile* esula dal livello di questo volume. In breve, sono misurabili le funzioni tali che l'immagine inversa di un intervallo aperto è un evento. La quasi totalità delle funzioni che si incontrano nella teoria e nelle applicazioni sono misurabili, perché le funzioni non-misurabili sono delle costruzioni matematiche piuttosto atipiche.

o Per ogni $E \subset \Omega$, la v.c. X assume un valore reale x. Quindi, tramite la v.c. si crea una corrispondenza tra il dominio Ω degli eventi e il codominio $\mathbb{R}$ dell'insieme dei numeri reali (equivalentemente, i punti di una retta), come mostra lo schema della Figura 9.1. L'insieme dei valori che una v.c. X può assumere con probabilità positiva, in una prova specifica, si chiamerà anche ***supporto*** della v.c. X.

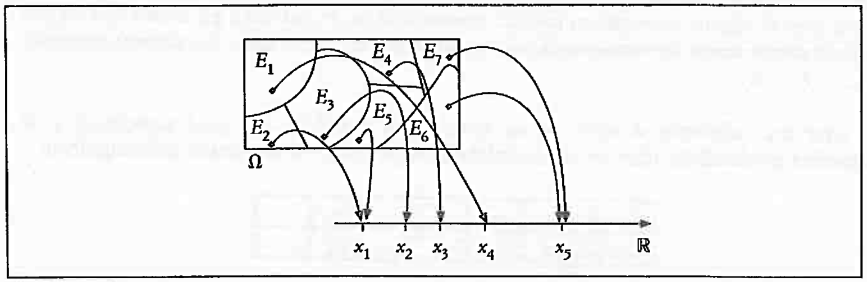

FIGURA 9.1. Corrispondenza tra eventi dello spazio campione e numeri reali.

Spesso, una v.c. non è altro che una traduzione numerica immediata degli eventi elementari, ma a seconda dei problemi specifici può essere anche una funzione più complicata definita sugli eventi di Ω.

Esempio 9.1

o Nella prova «lancio di un dado» che genera gli eventi elementari di $\Omega = \{E_1, E_2, \ldots, E_6\}$, è immediato definire la v.c. X che associa ad ogni faccia il «numero» x_i dei punti che essa presenta. In tal caso: $x_i = i$, $i = 1, 2, \ldots, 6$.

Ma è anche ben definita la v.c. Y =«numero dei punti sulla faccia che tocca il tavolo» che assume valori: $y_i = 7 - i$, $i = 1, 2, \ldots, 6$, oppure la v.c. Z = «il doppio dei punti che presenta la faccia di un dado meno il numero 7» che assume valori: $z_i = 2i - 7$, $i = 1, 2, \ldots, 6$. Infine, è anche ben definita la v.c. W = «la differenza fra 3 e il numero dei punti, espressa in valore assoluto» che assume i valori: $w_i = |3 - i|$, $i = 1, 2, \ldots, 6$; etc.

La tabella seguente, in corrispondenza dei 6 eventi elementari definiti sullo spazio Ω, mostra i valori reali che assumono le v.c. prima definite.

Eventi E_i	x_i	y_i	z_i	w_i
E_1	1	6	−5	2
E_2	2	5	−3	1
E_3	3	4	−1	0
E_4	4	3	1	1
E_5	5	2	3	2
E_6	6	1	5	3

o Analogamente, nella prova «lancio di due dadi», si possono – per esempio – definire le seguenti v.c.

S = «la somma dei punti che si presentano sulle due facce»;
T = «la differenza in valore assoluto tra i punti sulle due facce»;
U = «il valore massimo tra il numero dei punti sulle due facce».

□

L'esempio precedente evidenzia come la corrispondenza tra gli elementi di Ω e i numeri reali è univoca, ma non necessariamente biunivoca. Infatti, essendo una funzione degli eventi, ad ogni evento corrisponde un solo valore reale definito tramite la corrispondenza tra elementi di Ω e numeri reali; ma ciò non implica che il valore assunto dalla v.c. X derivi da un unico evento E.

Esempio 9.2

Nell'esempio 9.1, la v.c. X induce una corrispondenza biunivoca tra eventi e numeri reali perché vi è corrispondenza biunivoca tra le 6 facce di un dado ed i primi 6 numeri interi.

Tuttavia, ciò non avviene per le v.c. S, T, U. Per esempio, per la v.c. S, si ha che i seguenti 3 eventi distinti:

«faccia con 3 punti sul 1° dado e faccia con 1 punto sul 2° dado»;
«faccia con 2 punti sul 1° dado e faccia con 2 punti sul 2° dado»;
«faccia con 1 punto sul 1° dado e faccia con 3 punti sul 2° dado»,

sono tali che la v.c. S = «somma dei punti sulle due facce» assuma *sempre* il valore 4.

□

Dato uno *spazio di probabilità* $(\Omega, \mathcal{A}, Pr(.))$, dove $\mathcal{A}$ è la famiglia di tutti gli eventi generabili da Ω con un numero finito o numerabile di negazioni ed unioni (cioè $\mathcal{A}$ è la σ-algebra, o algebra di Boole completa, generata da Ω), sorge il problema di assegnare una probabilità non più sugli eventi di Ω ma sullo spazio generato dalla funzione misurabile, cioè sui numeri reali, ovvero sui valori assunti dalla v.c. in corrispondenza dei singoli eventi.

o La soluzione completa e formalmente dettagliata del problema supera il livello di questo testo, ed è comunque agevole solo se lo spazio Ω è discreto (finito o numerabile), perché in tale caso è sempre possibile elencare tutti i possibili eventi e le rispettive probabilità, associando le rispettive probabilità all'elenco dei valori reali corrispondenti agli eventi. Si dimostra, comunque, che tale corrispondenza è lecita anche per funzioni misurabili sugli eventi e quindi per tutte le v.c. di cui ci occuperemo nel seguito.

L'esempio seguente mostra il passaggio dallo spazio di probabilità ad un insieme reale per uno spazio di probabilità discreto e per una prova piuttosto semplice.

Esempio 9.3

Nella prova «lancio di 3 monete ben equilibrate», lo spazio Ω consiste degli 8 eventi elementari seguenti, ove abbiamo indicato per semplicità con TTT l'evento $(T_1 \cap T_2 \cap T_3)$, e similmente per gli altri.

Poiché gli 8 eventi elencati costituiscono una partizione equiprobabile di Ω, ciascuno di tali eventi elementari possiede probabilità 1/8.

Definiamo, quindi, la v.c. X = «numero di facce Testa» che si presentano in un determinato lancio. Tale funzione trasforma, ordinatamente, gli 8 eventi elementari di Ω nei numeri reali $3, 2, 2, 2, 1, 1, 1, 0$, come indicato nella tabella. Allora, l'insieme dei numeri reali che tale v.c. X può assumere è $\{0, 1, 2, 3\}$, e si chiama il *supporto* di X.

Eventi	TTT	TTC	TCT	CTT	TCC	CTC	CCT	CCC
Valori di X	3	2	2	2	1	1	1	0

Appare, allora, immediato attribuire ai valori di X una adeguata misura di probabilità, tramite l'assegnazione di probabilità originaria valida per gli eventi appartenenti ad Ω, secondo il ragionamento che esplicitiamo, in dettaglio, qui di seguito:

$$
\begin{aligned}
Pr(X=0) &= Pr(CCC) = \frac{1}{8};\\
Pr(X=1) &= Pr(TCC \cup CTC \cup CCT) = Pr(TCC) + Pr(CTC) + Pr(CCT) = \frac{3}{8};\\
Pr(X=2) &= Pr(TTC \cup TCT \cup CTT) = Pr(TTC) + Pr(TCT) + Pr(CTT) = \frac{3}{8};\\
Pr(X=3) &= Pr(TTT) = \frac{1}{8}.
\end{aligned}
$$

Si noti che lo spazio Ω consta di 8 eventi elementari equiprobabili mentre la v.c. X ha generato un supporto discreto e finito che consta solo di 4 valori distinti con probabilità:

Valori di X	0	1	2	3
Probabilità	1/8	3/8	3/8	1/8

Come vedremo, tale tabella costituisce la *distribuzione di probabilità della v.c. X* e consente di calcolare la probabilità di qualsiasi evento appartenente all'algebra di Boole $\mathcal{A}$ generata da Ω. Per esempio,

$$Pr(X<3) = Pr(X=0) + Pr(X=1) + Pr(X=2) = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} = \frac{7}{8}.$$

Similmente,

$$Pr(X \geq 2) = Pr(X=2) + Pr(X=3) = \frac{3}{8} + \frac{1}{8} = \frac{4}{8} = \frac{1}{2}.$$

□

9.3 Variabili casuali discrete

Una ***v.c. discreta*** è una corrispondenza tra gli eventi di $(\Omega, \mathcal{A}, Pr(.))$ ed un insieme discreto (finito o numerabile) di numeri reali.

o In quanto tale, una v.c. discreta è associata ai giochi, alle estrazioni di palline da urne, alla scelta di oggetti numerati, al numero di prove occorrenti per ottenere un certo risultato, etc.

Ma una v.c. discreta si può anche ottenere da uno spazio campione i cui eventi costituiscono una infinità non numerabile, mediante la tecnica della «discretizzazione».

Per esempio, nella prova «durata del tragitto in autostrada Roma-Napoli», se l'obiettivo è quello di giungere puntuali ad un appuntamento, può interessare solo se la durata supera 2 ore oppure no. In tal caso, all'evento E_x = «durata del percorso pari ad x» si fa corrispondere la v.c. D che vale 1 se la durata del percorso è inferiore a due ore, e vale 0 se la durata è uguale o supera le 2 ore. In tal caso, a livello di prova originaria, Ω contiene una infinità non numerabile di eventi (cioè tutte le possibili durate del tragitto) ma la v.c. è discreta perché assume solo i 2 possibili valori determinati dall'interesse che si ha per tale esperimento.

o In pratica, una v.c. discreta può assumere un numero finito di valori reali (e, in tal caso, gli eventi elementari da cui deriva costituiscono l'elenco finito dei possibili risultati di una prova) oppure un numero infinito *numerabile* (e, in tal caso, gli eventi elementari possono essere messi in corrispondenza con gli elementi della serie dei numeri naturali: $1, 2, \ldots, i, \ldots$).

Una v.c. discreta è nota se si conoscono i valori che può assumere e le rispettive probabilità, cioè se sono ordinatamente note le seguenti informazioni:

Valori di X	x_1	x_2	...	x_i	...
Probabilità	p_1	p_2	...	p_i	...

La tabella precedente viene definita ***distribuzione della probabilità della v.c. discreta*** X e rappresenta un insieme di informazioni necessarie e sufficienti per calcolare la probabilità di qualsiasi evento concernente la v.c. X. Ovviamente, tali probabilità devono rispettare i postulati del Calcolo delle probabilità (paragrafo 8.9).

Pertanto, condizione necessaria e sufficiente affinché una v.c. discreta X sia *ben definita* è che tali probabilità soddisfino le seguenti condizioni:

$$i)\ p_i \geq 0, \forall i = 1, 2, \ldots; \qquad ii)\ \sum_{i=1}^{\infty} p_i = 1.$$

Si noti che nessun vincolo viene imposto sui valori reali che una v.c. può assumere.

La *rappresentazione grafica* più naturale di una v.c. discreta è quella di porre in ascissa i valori $(x_1, x_2, \ldots)$ che la v.c. può assumere ed in ordinata le corrispondenti probabilità $(p_1, p_2, \ldots)$, secondo la Figura seguente:

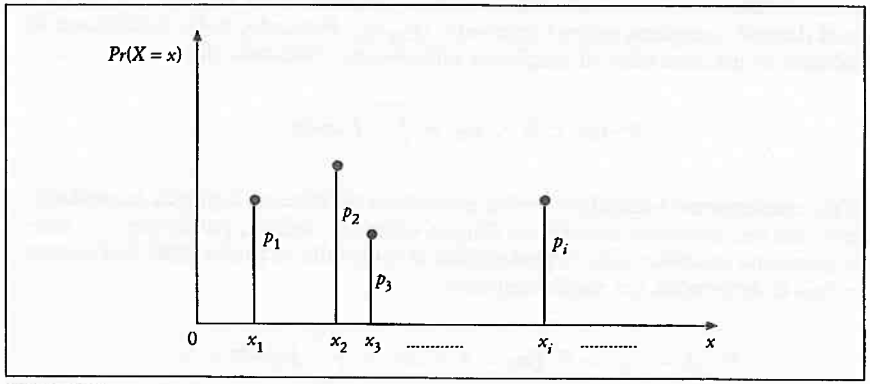

FIGURA 9.2. Distribuzione di probabilit di una variabile casuale discreta.

Nel precedente esempio 9.3 abbiamo sviluppato un caso di v.c. discreta finita, mentre nell'esempio seguente presentiamo un caso di v.c. discreta numerabile.

Esempio 9.4

Un'urna contiene un numero finito di palline numerate in modo che risulti pari a $\theta \in (0,1)$ la proporzione delle palline recanti un *numero pari*. Per esempio, nella tombola (che contiene bussolotti numerati da 1 a 90) sarà $\theta = 1/2$.

La prova consiste nell'estrarre con ripetizione palline sino a che non si verifichi l'evento $E =$ «la pallina estratta presenta un numero pari».

Su tale prova, definiamo la v.c. $X =$ «numero di estrazioni occorrenti perché si verifichi E per la prima volta».

Ora, se la prima pallina estratta è pari si verifica immediatamente E alla prima estrazione per cui la v.c. assume il valore 1, cioè si verifica $(X = 1)$ con probabilità θ. Invece, se la prima pallina estratta è dispari e la seconda è pari allora (con ovvia simbologia) si verifica l'evento $\overline{E} \cap E$ e la v.c. assume il valore 2, cioè si verifica $(X = 2)$, con probabilità $\theta(1-\theta)$, perché le sottoprove sono indipendenti, e così via.

In generale, se nelle prime $i-1$ estrazioni si verifica una pallina dispari e poi alla i-esima estrazione si verifica una pallina pari, allora si verifica l'evento:

$$\underbrace{\overline{E} \cap \overline{E} \cap \cdots \cap \overline{E}}_{(i-1)\ volte} \cap E$$

e la v.c. assume il valore i, cioè si verifica $(X = i)$ con probabilità:

$$\underbrace{(1-\theta)(1-\theta)\ldots(1-\theta)}_{(i-1)\ volte}\,\theta = (1-\theta)^{i-1}\theta, \quad i = 1, 2, \ldots.$$

Pertanto, lo spazio delle probabilità derivato da questa prova è:

$$\Omega = \{E,\ \overline{E} \cap E,\ \overline{E} \cap \overline{E} \cap E,\ \ldots,\ \overline{E} \cap \cdots \cap \overline{E} \cap E,\ \ldots\}$$

A tali eventi, grazie alla v.c. X sopra definita, corrispondono ordinatamente i numeri interi positivi (che sono una infinità numerabile, per cui la v.c. X è discreta numerabile) e le corrispondenti probabilità espresse nella tabella seguente:

Valori di X	1	2	3	...	i	...
Probabilità	θ	$(1-\theta)\theta$	$(1-\theta)^2\theta$	...	$(1-\theta)^{i-1}\theta$	...

In questo esempio, l'intera distribuzione di probabilità della v.c. si può sintetizzare in una sola espressione:

$$Pr(X = i) = (1-\theta)^{i-1}\theta, \qquad i = 1, 2, \ldots$$

perché essa, al variare di $i = 1, 2, \ldots$, esprime sia i valori che la v.c. assume che le corrispondenti probabilità; il che equivale a conoscere compiutamente la v.c. discreta X.

La distribuzione di probabilità della v.c. X è ben definita perché ciascuna delle probabilità è maggiore o uguale a 0 e la somma (in questo caso, la serie trattandosi di una somma con infiniti addendi) vale esattamente 1, come è agevole dimostrare sfruttando le proprietà della serie geometrica.

È possibile allora calcolare la probabilità di qualsiasi evento di interesse; per esempio, si può calcolare la probabilità che la prova termini dopo 4 estrazioni:

$$\begin{aligned} Pr(X > 4) &= 1 - Pr(X \leq 4) \\ &= 1 - [Pr(X = 1) + Pr(X = 2) + Pr(X = 3) + Pr(X = 4)] \\ &= 1 - \left[\theta + (1-\theta)\theta + (1-\theta)^2\theta + (1-\theta)^3\theta\right] \\ &= 1 - 4\theta + 6\theta^2 - 4\theta^3 + \theta^4 = (1-\theta)^4. \end{aligned}$$

Il risultato è ovvio se si pensa che l'evento $(X > 4)$ si verifica se e solo se non esce un numero pari per le prime 4 estrazioni; ebbene, tale evento, per l'indipendenza delle sottoprove, possiede probabilità:

$$Pr(X > 4) = Pr\left(\overline{E} \cap \overline{E} \cap \overline{E} \cap \overline{E}\right) = (1-\theta)(1-\theta)(1-\theta)(1-\theta) = (1-\theta)^4.$$

□

9.4 Variabili casuali continue

Abbiamo già anticipato che una v.c. è continua se può assumere tutti i valori compresi in un (qualsiasi) intervallo reale. Formalmente, una ***v.c. continua*** X è una funzione misurabile e a valori reali che assegna ad ogni evento $E \subset \Omega$ di uno spazio di probabilità continuo un numero reale $x \in \mathbb{R}$.

o Le v.c. continue presentano una maggiore complessità analitica poiché per esse non è possibile elencare tutti i valori che la v.c. assume, essendo questi una infinità non-numerabile. Occorre, quindi, assegnare la probabilità ad intervalli sulla retta e derivare, poi, da tale assegnazione tutte le probabilità degli eventi che interessano.

In prima istanza, tale assegnazione sembrerà astratta e poco collegata alla specifica prova ma, più oltre, si vedrà come l'introduzione di una probabilità per gli intervalli è strettamente derivata dalla natura e dalla specificità della prova.

Per ragioni didattiche, discuteremo di alcuni concetti intuitivi senza il pur necessario rigore formale allo scopo di pervenire ad una esposizione essenziale delle nozioni occorrenti per il prosieguo del corso.

*Una **v.c. continua** X è nota se, per ogni x_0 reale e prefissato, è nota la probabilità che tale v.c. assuma un valore in un intervallino di ampiezza infinitesimo rispetto a x_0* mediante la relazione seguente:

$$P_r(x_0 < X \leq x_0 + dx) = f(x_0)dx\,,$$

dove $f(x)$ è una funzione a valori reali, detta ***funzione di densità*** della v.c. continua X. Spesso, scriveremo: $X \sim f(x)$ per indicare che la v.c. X possiede $f(x)$ come propria funzione di densità.

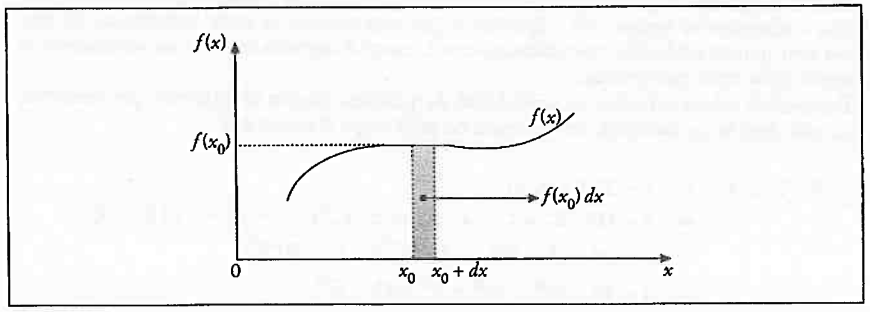

FIGURA 9.3. Probabilit di un intervallo infinitesimo e funzione di densit .

Come la Figura 9.3 illustra, mediante tale definizione la probabilità che la v.c. X assuma valori nell'intervallino raffigurato è uguale a $f(x)\,dx$ perché essa è uguale all'area del rettangolo la cui base misura dx e la cui altezza è $f(x_0)$, perché per la continuità deve essere: $f(x_0) = f(x_0 + dx)$. Si osservi, comunque, che un intervallino di ampiezza infinitesima non potrebbe essere fisicamente disegnato, per cui la illustrazione ha solo una finalità didattica.

Quindi, la probabilità di un intervallino infinitesimo è l'area che è sottesa dalla funzione di densità al di sopra dell'intervallino di riferimento e delimitata dall'asse delle ascisse.

Da ciò discende che, se si desidera calcolare la probabilità che la v.c. continua X assuma un valore in un intervallo di ampiezza finita, per esempio $(x_1, x_2]$, cioè se si desidera calcolare $P_r(x_1 < X \leq x_2)$, allora si può calcolare l'area di tale intervallo «sommando nel continuo» tutti gli infiniti intervallini di ampiezza infinitesima che sono compresi tra x_1 e x_2, e cioè:

$$(x_1, x_1 + dx],\ (x_1 + dx, x_1 + 2dx],\ (x_1 + 2dx, x_1 + 3dx],\ \ldots,\ (x_2 - dx, x_2]$$

Ebbene, tale «somma nel continuo» equivale in matematica alla integrazione della funzione $f(x)$ tra gli estremi x_1 e x_2 dell'intervallo in questione; tale integrale definito è esattamente l'area sottesa dall'asse delle ascisse e dalla funzione di densità racchiusa entro l'intervallo $(x_1, x_2]$. Pertanto, dalla definizione di probabilità di un intervallo di ampiezza infinitesima, discende che:

$$P_r(x_1 < X \leq x_2) = \int_{x_1}^{x_2} f(x)dx.$$

Una conseguenza inaspettata delle precedenti definizioni riguarda la probabilità che una v.c. continua assuma un singolo valore x_0; infatti, poiché per v.c. continue possiamo calcolare solo le probabilità di intervalli, la probabilità dell'evento $(X = x_0)$ si determina nel modo seguente:

$$P_r(X = x_0) = P_r(x_0 < X \leq x_0) = \int_{x_0}^{x_0} f(x)dx = 0.$$

Le precedenti espressioni mostrano che, per una v.c. continua X, la probabilità che X assuma valori in qualsiasi intervallo finito è sempre misurata dall'area sottesa dalla funzione di densità sull'intervallo medesimo. Ne discende, allora, che la probabilità di intervalli di pari ampiezza sarà tanto più elevata quanto più la funzione di densità assumerà ivi valori maggiori.

Quindi, *la struttura di una prova e la probabilità degli eventi da essa generati determinano la funzione di densità*, e viceversa.

o Per esempio, una funzione di densità costantemente decrescente da 0 a $+\infty$ deriva da una prova che attribuisce agli intervalli $[0, 100]$, $[100, 200]$, $\ldots$, delle probabilità sempre più piccole. Per contro, una funzione di densità costante su tali intervalli deriva da una prova nella quale gli eventi $(0 < X \leq 100), (100 < X \leq 200), \ldots$, hanno tutti la stessa probabilità. Quindi, se si deve studiare la v.c. «durata di funzionamento di una lampadina elettrica» sembra più ragionevole ipotizzare una funzione di densità del primo tipo.

Condizioni necessarie e sufficienti perché una v.c. continua sia ben definita sono:

$$i)\ f(x) \geq 0; \qquad ii) \int_{-\infty}^{+\infty} f(x)dx = 1.$$

o Nei casi concreti, avviene spesso che la v.c. continua assuma valori realizzabili solo su un sottoinsieme più limitato dell'intera retta reale, e tale insieme di valori reali costituisce quindi l'effettivo *supporto* di tale v.c.

Per esempio, la durata di una corsa può assumere solo valori positivi; la direzione prevalente del vento espressa in radianti sarà solo un angolo compreso tra $[-\pi, +\pi]$; l'altezza in cm di uno studente universitario, casualmente prescelto, sarà un numero reale incluso in $[140, 210]$, e così via.

Tuttavia, si può anche ipotizzare che *tutte* le v.c. continue siano sempre definite sull'intero asse reale, attribuendo funzione di densità pari a 0 al di fuori dello specifico supporto.

Così, se una particolare v.c. continua X è definita sull'intervallo (a, b) con funzione di densità $f(x) > 0$, essa equivale a tutti gli effetti alla v.c. continua X^* che possiede funzione di densità $g(x)$ definita da:

$$g(x) = \begin{cases} f(x), & a \le x \le b; \\ 0, & \text{altrove.} \end{cases}$$

Per tale motivo, nel seguito, parlando di una v.c. continua supporremo che essa sia sempre definita sull'intera retta reale anche se poi, di volta in volta, si individua lo specifico intervallo effettivo su cui essa assume una funzione di densità diversa da 0.

o Infine, si noti che *la funzione di densità non è una probabilità*, ma è solo proporzionale alla probabilità che la v.c. assuma valori in un intervallo infinitesimo. Quindi, la funzione di densità potrebbe anche essere superiore ad 1, mentre l'area sottesa su qualsiasi sotto-intervallino deve essere strettamente minore o uguale ad 1.

9.5 La funzione di ripartizione di una variabile casuale

La funzione di ripartizione costituisce l'elemento unificatore della teoria delle v.c. discrete e continue perché essa è definita nello stesso modo per entrambi i tipi di v.c. anche se il calcolo si sviluppa, poi, tramite una sommatoria (per le v.c. discrete) oppure un integrale definito (per le v.c. continue).

Formalmente, la ***funzione di ripartizione*** $F(x_0)$ di una v.c. X calcolata nel punto x_0 è definita dalle relazioni seguenti:

$$F(x_0) = P_r(X \le x_0) = \begin{cases} \displaystyle\sum_{x \le x_0} p_i, & \text{se X è una v.c. discreta;} \\ \displaystyle\int_{-\infty}^{x_0} f(w)dw, & \text{se X è una v.c. continua.} \end{cases}$$

Dalla definizione, discendono le seguenti proprietà di una funzione di ripartizione:

i) $F(x)$ è non-decrescente, cioè: $x_1 < x_2 \Rightarrow F(x_1) \le F(x_2)$;

ii) $\lim_{x \to -\infty} F(x) = 0$; $\lim_{x \to +\infty} F(x) = 1$;

iii) $F(x)$ è continua da destra, cioè: $\lim_{x \to x_0^+} F(x) = F(x_0)$.

o Dalla i) e dalla ii) si deduce: $0 \le F(x) \le 1$, $\forall x \in \mathbb{R}$, in accordo con il fatto che $F(x)$ è, in effetti, una probabilità.

o La funzione $F(x)$ è non-decrescente perché sia nel caso discreto (sommando delle probabilità) che nel caso continuo (integrando la funzione di densità) essa si accresce o al più rimane costante al variare di x tra $-\infty$ (quando la funzione di ripartizione vale 0) e $+\infty$ (quando vale 1).

o Infine, la continuità a destra (cioè *da destra*, ovvero per x tendente ad x_0 da valori più grandi di x_0) deriva dall'aver definito $F(x)$ come la probabilità dell'evento $(X \le x_0)$.

o Alcuni Autori definiscono: $F(x) = P_r(X < x)$, il che implica la continuità da sinistra. Nel testo, ci siamo uniformati alla letteratura internazionale prevalente.

o Nel seguito, se la v.c. X possiede funzione di ripartizione $F(x)$, scriveremo: $X \sim F(x)$.

È fondamentale rendersi conto che la conoscenza della funzione $F(x)$ equivale alla conoscenza dell'intera struttura probabilistica della v.c. X, come ora espliciteremo distinguendo il caso continuo da quello discreto.

• *Se X è una v.c. continua*, $F(x)$ è una funzione assolutamente continua su $\mathbb{R}$ (con l'eccezione di eventi di probabilità 0). Ora, ogni evento può essere sempre ricondotto alla unione, intersezione, negazione (in numero finito o numerabile) di intervalli del tipo $(-\infty, x]$ e la funzione di ripartizione è definita precisamente per questi intervalli.

Pertanto, in generale, si ha:

$$P_r(x_1 \le X \le x_2) = P_r(X \le x_2) - P_r(X \le x_1) = F(x_2) - F(x_1), \forall x_1 \le x_2.$$

Da tale relazione, si deducono tutte le altre probabilità particolarizzando opportunamente x_1 e/o x_2. Per esempio, per calcolare la probabilità che X sia non inferiore a x_1 (cioè maggiore o uguale a x_1) basta porre $x_2 = +\infty$ nella precedente relazione:

$$P_r(X \ge x_1) = P_r(x_1 \le X < +\infty) = P_r(X < +\infty) - P_r(X \le x_1) = 1 - F(x_1), \ \forall x_1 < \infty.$$

Inoltre, per una v.c. continua, essendo nulla la probabilità di un singolo valore, le seguenti probabilità coincidono:

$$P_r(x_1 < X \le x_2) \equiv P_r(x_1 \le X < x_2) \equiv P_r(x_1 \le X \le x_2) \equiv P_r(x_1 < X < x_2);$$

pertanto, quando si calcolano probabilità relative a v.c. continue è irrilevante la inclusione o la esclusione dell'estremo (degli estremi) di un intervallo.

o Per v.c. continue, il legame tra la funzione di ripartizione $F(x)$ e la funzione di densità $f(x)$ è espresso dalle relazioni seguenti:

$$F(x) = \int_{-\infty}^{x} f(w)\, dw; \qquad f(x) = \frac{d}{dx} F(x),$$

per cui dalla conoscenza dell'una si può passare a quella dell'altra.

• *Se X è una v.c. discreta*, $F(x)$ è una funzione a gradini che è costante su intervalli dell'asse reale e cresce un numero discreto (finito o numerabile) di volte in corrispondenza dei valori $x_1, x_2, \ldots, x_i, \ldots$ assunti dalla v.c. discreta: il «gradino» è esattamente pari alle probabilità $p_1, p_2, \ldots, p_i, \ldots$ con le quali la v.c. discreta assume quei valori.

Si noti che la funzione di ripartizione è ben definita su tutto l'asse reale, anche se la v.c. discreta può assumere con probabilità positiva solo un numero distinto e ben specificato

di valori. Infatti, la funzione di ripartizione calcolata in x_0 non riguarda solo il valore x_0, ma rappresenta la probabilità corrispondente a tutti i valori *inferiori od uguali a* x_0.

o Una differenza sostanziale tra la funzione di ripartizione di una v.c. discreta e quella di una v.c. continua si manifesta nel calcolo della probabilità di un valore singolo. Infatti, come abbiamo visto, tale probabilità è in ogni caso pari a:

$$Pr(X = x_0) = Pr(x_0 < X \leq x_0) = F(x_0) - F(x_0^-)$$

per ogni x_0 reale, dove $F(x_0^-)$ è il limite di $F(x)$ per x che tende a x_0 da sinistra.

Ora, *se X è una v.c. continua* la funzione di ripartizione $F(x)$ sarà continua per cui: $F(x_0^-) = F(x_0)$ e, quindi, $Pr(X = x_0) = 0$.

Invece, *se X è una v.c. discreta*, essendo $F(x)$ una funzione a gradini, l'espressione precedente non sarà sempre nulla perché nei punti di discontinuità (che coincidono con il supporto della v.c. discreta) il limite sinistro $F(x_0^-)$ non coincide necessariamente con $F(x_0)$. In tali punti, la differenza $F(x_0) - F(x_0^-)$ è proprio la probabilità che la v.c. assuma il valore x_0, come sarà evidente nell'esempio 9.5. Invece, in tutti gli altri punti non appartenenti al supporto della v.c. X, tale differenza è identicamente nulla.

Riassumiamo, nel riquadro seguente, il calcolo delle probabilità relative ad un intervallo reale, per i quattro casi possibili:

$$\begin{aligned}
Pr(x_1 < X \leq x_2) &= F(x_2) - F(x_1) \\
Pr(x_1 \leq X \leq x_2) &= F(x_2) - F(x_1) + Pr(X = x_1) \\
Pr(x_1 \leq X < x_2) &= F(x_2) - F(x_1) + Pr(X = x_1) - Pr(X = x_2) \\
Pr(x_1 < X < x_2) &= F(x_2) - F(x_1) - Pr(X = x_2)
\end{aligned}$$

Tale schema è valido per qualsiasi v.c. X (discreta o continua), con l'intesa che nel caso di una v.c. continua occorre ricordare che: $Pr(X = x_1) = Pr(X = x_2) \equiv 0$.

Sviluppiamo, ora, due esempi per una v.c. discreta e una v.c. continua, rispettivamente, nei quali sono discussi il calcolo, la rappresentazione e la utilizzazione della funzione di ripartizione.

Esempio 9.5

Sia X la v.c. discreta associata al lancio di un dado ben equilibrato, il cui supporto è costituito dai valori: $x = 1, 2, \ldots, 6$, tutti con probabilità $1/6$, per cui la sua distribuzione di probabilità è riassunta nella tabella seguente:

Valori della v.c. X	1	2	3	4	5	6	*Totale*
Probabilità	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

Esplicitamente, la funzione di ripartizione di tale v.c. è così definita:

$$F(x) = \begin{cases} 0, & -\infty < x < 1 \\ \frac{1}{6}, & 1 \leq x < 2 \\ \frac{2}{6}, & 2 \leq x < 3 \\ \frac{3}{6}, & 3 \leq x < 4 \\ \frac{4}{6}, & 4 \leq x < 5 \\ \frac{5}{6}, & 5 \leq x < 6 \\ 1, & 6 \leq x < +\infty \end{cases}$$

Come si vede dalla Figura 9.4, essa è una funzione «a gradini» la cui altezza è pari alle probabilità di assumere i valori del supporto della v.c. X; questi ultimi costituiscono punti di discontinuità per la funzione di ripartizione.

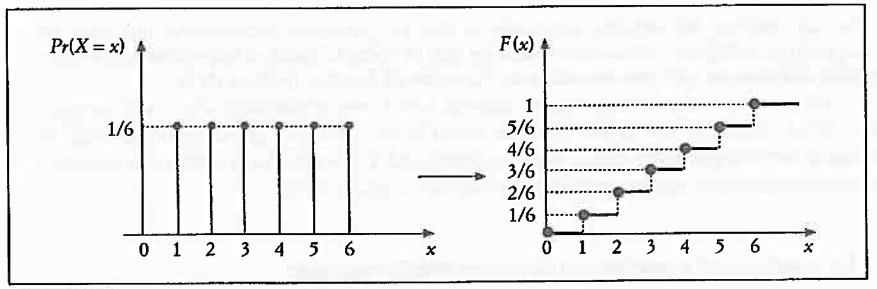

FIGURA 9.4. Distribuzione di probabilit e funzione di ripartizione della variabile casuale discreta associata alla prova «Lancio di un dado simmetrico».

Si noti che $F(x)$ è definita per ogni x reale; così, $F(-0.37) = 0$ perché non può mai verificarsi l'evento $(X \leq -0.37)$. Invece, $F(3.14) = 3/6$ perché $F(x) = \frac{3}{6}$, per $3 \leq x < 4$ e l'evento $(X \leq 3.14)$ si verifica se e solo se si verifica uno degli eventi $(X = 1)$ oppure $(X = 2)$ oppure $(X = 3)$; infine, $F(12) = 1$ perché è certo che si verificherà l'evento $(X \leq 12)$, in quanto esso include tutti i valori possibili per la v.c. X.

Inoltre, è agevole convincersi che:

$$F(3) = F(3.14) = F(3.57) = F(3.76) = F(3.90) = F(3.999) = F(3.999999) = \frac{3}{6};$$

mentre:

$$F(4) = F(4.0001) = F(4.21) = F(4.79) = F(4.90) = F(4.999999) = \frac{4}{6};$$

e così via. Tali risultati derivano dalla proprietà della continuità da destra per $F(x)$ da cui deriva, per esempio, che $F(3+) = F(3)$. Il grafico evidenzia come, in generale, si abbia: $F(x+) = F(x)$, per ogni x reale.

□

Esempio 9.6

Supponiamo che una v.c. continua X possieda funzione di densità che è sempre nulla eccetto per $x \in [0,1]$, dove cresce proporzionalmente con x. Quindi, $f(x) \equiv 0$, per $x \notin [0,1]$, mentre: $f(x) = cx$, per $x \in [0,1]$, essendo c una costante reale.

Affinché tale v.c. sia ben definita, occorre che la funzione di densità sia non-negativa e l'area complessiva da essa sottesa sia pari a 1. Ciò impone, anzitutto, $c > 0$ ed inoltre:

$$1 = \int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{0} (0)dx + \int_{0}^{1} (cx)dx + \int_{1}^{\infty} (0)dx = \frac{c}{2} \iff c = 2.$$

Quindi, la funzione di densità è univocamente specificata ed è così definita:

$$f(x) = \begin{cases} 2x & \text{per ogni } 0 \leq x \leq 1; \\ 0, & \text{altrove.} \end{cases}$$

Per ottenere la funzione di ripartizione, occorre tener presente che – per una v.c. continua – $F(x)$ rappresenta l'area sottesa dalla funzione di densità per valori non superiori a x; quindi, essa è nulla sino ad $x = 0$; vale 1 per $x \geq 1$. Nel supporto specifico dove essa non è nulla, cioè per ogni $x \in [0,1]$, si ha:

$$F(x) = \int_{-\infty}^{x} f(w)dw = \int_{-\infty}^{0} (0)dw + \int_{0}^{x} (2w)dw = x^2.$$

Pertanto, la funzione di ripartizione di tale v.c. è così definita:

$$F(x) = \begin{cases} 0 & -\infty < x \leq 0; \\ x^2, & 0 < x \leq 1; \\ 1, & 1 < x < +\infty. \end{cases}$$

Tramite $F(x)$ – disegnata nella Figura 9.5 – si possono calcolare le probabilità di qualsiasi evento che riguarda la v.c. X; per esempio:

$$\begin{aligned} Pr(X \leq 0.4) &= F(0.4) = (0.4)^2 = 0.16; \\ Pr(X > 0.5) &= 1 - F(0.5) = 1 - (0.5)^2 = 0.75; \\ Pr(0.1 < X \leq 0.4) &= F(0.4) - F(0.1) = (0.4)^2 - (0.1)^2 = 0.15; \\ Pr((X < 0.7) \cup (X \geq 0.3)) &= Pr(X < 0.7) + Pr(X \geq 0.3) - Pr((X < 0.7) \cap (X \geq 0.3)) \\ &= (0.7)^2 + (1 - 0.3^2) - Pr(0.3 \leq X < 0.7) \\ &= 0.49 + 1 - 0.09 - (0.49 - 0.09) = 1. \end{aligned}$$

In effetti, l'ultimo evento consiste dell'intero supporto per tale v.c., e quindi la sua probabilità vale 1 (essendo quella dell'evento certo).

Si osservi, infine, che data la semplicità della funzione di densità (una linea retta di equazione $f(x) = 2x$, limitatamente all'intervallo $[0,1]$), tutti i risultati precedenti potevano essere ricavati senza ricorrere al calcolo integrale, ma calcolando l'area di figure geometriche elementari.

Per esempio, per ogni $x \in [0,1]$, $F(x)$ è l'area del triangolo rettangolo di vertici $(0,0)$, $(x,0)$, $(x,2x)$, ed è quindi uguale a: $F(x) = x(2x)/2 = x^2$.

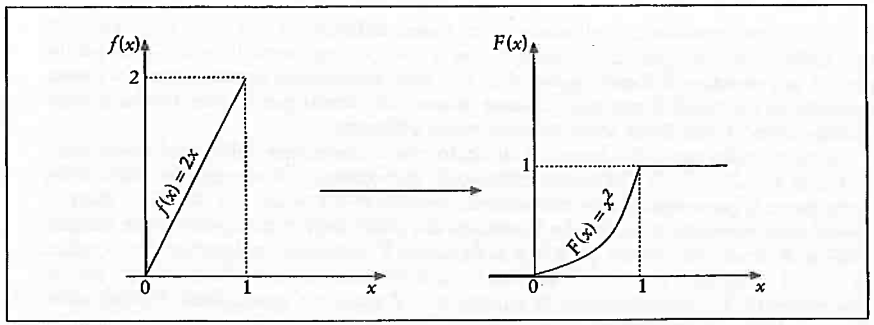

FIGURA 9.5. Funzione di densit e funzione di ripartizione della variabile casuale continua definita nell'esempio 9.6.

Similmente, la probabilità di un qualsiasi intervallo $[x_0, x_1] \subset [0,1]$, è l'area del trapezio rettangolo la cui base minore passa per x_0 (ed è lunga $2x_0$), la base maggiore passa per x_1 (ed è lunga $2x_1$) e l'altezza è il segmento di ascissa compreso tra x_0 e x_1 (ed è lungo $x_1 - x_0$). Quindi, tale area è:

$$Pr(x_0 \leq X \leq x_1) = (2x_0 + 2x_1)(x_1 - x_0)/2 = (x_1)^2 - (x_0)^2.$$

□

Abbiamo già mostrato come la conoscenza della funzione di ripartizione sia necessaria e sufficiente per specificare completamente una v.c. A tale riguardo, sono opportune alcune ulteriori specificazioni.

• Date due v.c. X e Y diremo che esse sono ***somiglianti*** se possiedono la stessa funzione di ripartizione. Quindi, X e Y avranno le stesse caratteristiche probabilistiche, lo stesso supporto, la medesima probabilità per gli stessi eventi e, come vedremo, gli stessi indicatori sintetici; tuttavia, se X e Y sono v.c. somiglianti non significa che assumeranno lo stesso valore in una specificata prova.

• Se, invece, per due v.c. X e Y avviene che: $Pr(X = Y) = 1$, diremo che X e Y sono v.c. ***coincidenti*** (talvolta, si dice anche *uguali*).

È bene sottolineare che somiglianza e coincidenza tra v.c. sono concetti completamente distinti, come è confermato dall'esempio seguente.

Esempio 9.7

o Se lanciamo due monete in modo indipendente e facciamo corrispondere alla faccia Testa il valore 1 e alla faccia Croce il valore 0, allora la v.c. X = «esito della prima moneta» e la v.c. Y = «esito della seconda moneta» sono somiglianti poiché hanno la stessa distribuzione di probabilità e, quindi, la stessa funzione di ripartizione. Ma, come è evidente, *questo non implica che lanciando le due monete esse presentino la stessa faccia*, cioè si abbia $X = Y$ con probabilità 1. Quindi, le due v.c. X e Y così definite sono somiglianti ma non coincidenti.

o Nella prova consistente nel lancio di un dado, indichiamo con X la v.c. «numero di punti della faccia che guarda in alto» e con Y la v.c. «numero di punti della faccia che tocca il pavimento». Ebbene, queste due v.c. sono somiglianti perché hanno la stessa distribuzione di probabilità ma non possono essere coincidenti per ragioni fisiche perché sicuramente nella stessa prova assumeranno valori differenti.

o Tuttavia, nella prova del lancio di un dado, con le medesime definizioni precedenti, la v.c. X e la v.c. $Z = 7 - Y$ (ottenuta sottraendo dal numero 7 il numero dei punti della faccia che tocca il pavimento) sono coincidenti, perché: $Pr(X = Z) = 1$. Infatti, i dadi in commercio sono costruiti in modo che la somma dei punti delle facce opposte vale sempre 7, per cui se X assumerà i valori $\{1, 2, 3, 4, 5, 6\}$ allora Y assumerà, ordinatamente, i valori $\{6, 5, 4, 3, 2, 1\}$ e, quindi, $Z = 7 - Y$ assumerà esattamente, come risultato di ogni prova, lo stesso valore di X con probabilità 1: quindi, X e Z sono v.c. coincidenti. Ovviamente, X e Z sono anche v.c. somiglianti. □

9.6 Valori medi di una variabile casuale

Abbiamo visto che una v.c. è nota se è nota la sua funzione di ripartizione, ovvero se sono noti i valori che assume e la distribuzione di probabilità (se è discreta) oppure la funzione di densità (se è continua). Tuttavia, per confrontare prove differenti, si rende spesso necessario ed utile il confronto tra alcuni indicatori sintetici costruiti sulla v.c. Evidentemente, ogni sintesi deve tener conto sia dei valori della v.c. che delle rispettive probabilità.

In questo paragrafo, dopo una discussione sui principi che giustificano tali sintesi, si introdurranno esclusivamente il *valore medio* e la *varianza* di una v.c., tramite i quali sarà poi introdotta la v.c. standardizzata; di questi indicatori si specificheranno anche le principali proprietà.

o Un modo intuitivo per sintetizzare una v.c. discreta è quello di moltiplicare i suoi valori per le rispettive probabilità e poi di sommarli: l'indice che ne risulterà sarà modesto se a valori piccoli si associano piccole probabilità e sarà elevato se a valori grandi si associano grandi probabilità; assumerà valori intermedi negli altri casi. La gran parte degli indicatori successivi obbedisce a questo criterio elementare di sintesi, ma è opportuno dare qualche giustificazione che deriva da una interpretazione probabilistica della prova.

o Supponiamo che in una prova $\mathcal{P}_1$ si vince 1 Euro se, lanciando una moneta ben equilibrata, si verifica Testa o Croce (cioè sempre) e in un'altra prova $\mathcal{P}_2$ si vincono 2 Euro se esce Testa e 0 Euro se esce Croce: come confrontiamo queste due situazioni? Evidentemente, supponiamo costante l'*attitudine al rischio* perché, altrimenti, qualcuno preferirà in ogni caso $\mathcal{P}_1$ a $\mathcal{P}_2$ («meglio l'uovo oggi che la gallina domani!»). Stiamo, quindi, confrontando due prove: nella prima il valore 1 ha probabilità 1 di verificarsi; nella seconda, il valore 2 e il valore 0 hanno entrambi probabilità 1/2 di verificarsi.

• Per un *frequentista* il confronto è semplice: per $\mathcal{P}_1$ egli incasserà sempre 1; per $\mathcal{P}_2$ incasserà, nel lungo periodo, 2 Euro metà delle volte e 0 per l'altra metà delle volte; quindi, la vincita attesa è 1 Euro per $\mathcal{P}_1$ e vale: $2\,(1/2) + 0\,(1/2) = 1$ per $\mathcal{P}_2$. Un frequentista dirà che le due prove sono equivalenti.

In generale, se incassa $x_1, x_2, \ldots$ in corrispondenza di eventi $E_1, E_2, \ldots$ che hanno probabilità $p_1, p_2, \ldots$ tali che $\sum p_i = 1$, allora il frequentista sa che, nel lungo periodo, ciascuna frequenza relativa (n_i/n) tenderà alla rispettiva probabilità p_i per cui egli incasserà, ripetendo molte volte lo stesso gioco:

$$x_1(n_1/n) + x_2(n_2/n) + \ldots \;\to\; x_1 p_1 + x_2 p_2 + \ldots .$$

Egli confronterà questa espressione con il valore analogo di un'altra prova o con quello di una prova ad esito certo che gli restituirà quel medesimo valore.

• Il *soggettivista* è disposto a pagare $Pr(E)$ se incassa 1 Euro quando si verifica E e, ovviamente, $x_E Pr(E)$ se incassa x_E Euro se si verifica E. Se, quindi, il soggettivista scommette $x_1 p_1$, $x_2 p_2, \ldots$ per ricevere, ordinatamente, $x_1, x_2, \ldots$ se si verificano gli eventi $E_1, E_2, \ldots$, allora (essendo gli eventi necessari perché: $p_1 + p_2 + \cdots = 1$) egli sarà disposto a scommettere $x_1 p_1 + x_2 p_2 + \ldots$ per ricevere 1 Euro se almeno uno degli eventi $E_1, E_2, \ldots$ si verifica.

o Come si vede, a prescindere dalla impostazione cui aderiscono, sia il frequentista che il soggettivista propongono il valore numerico: $x_1 p_1 + x_2 p_2 + \ldots$ per sintetizzare e confrontare una prova ad esito incerto. In un certo senso, questo numero rappresenta l'*equivalente certo di una prova incerta* i cui esiti $x_1, x_2, \ldots$ hanno probabilità $p_1, p_2, \ldots$ Tale valore numerico è l'ammontare da scommettere (con certezza) a fronte delle quantità $x_1, x_2, \ldots$ (incerte) che si incasseranno con probabilità $p_1, p_2, \ldots$

Data una v.c. X ben definita, si definisce il ***valore medio*** della v.c. X la seguente quantità:

$$\mathbb{E}(X) = \begin{cases} \sum_{i=1}^{\infty} x_i p_i & \text{se } X \text{ è una v.c. discreta;} \\ \int_{-\infty}^{+\infty} x f(x)\,dx, & \text{se } X \text{ è una v.c. continua;} \end{cases}$$

se e solo se il valore medio di $|X|$ esiste finito (questa precisazione si rende necessaria per utilizzare serie ed integrali definiti su insiemi che possono essere, eventualmente, di ampiezza infinita).

Il valore medio di X verrà sempre indicato con il simbolo $\mathbb{E}(X)$, che deriva dall'inglese *Expectation*=«aspettativa», ed esplicita il fatto che trattasi di una sintesi derivata dalla v.c. X.

In sostanza, il valore medio di una v.c. discreta X è la somma dei prodotti dei valori x_i che la v.c. X assume per le corrispondenti probabilità p_i. Analogamente, il valore medio di una v.c. continua X è l'integrale del prodotto dei valori ammissibili (x) per la funzione di densità $f(x)$, come mostra l'esempio seguente.

Esempio 9.8

o Per la v.c. discreta X di cui all'esempio 9.5, il valore medio è:

$$\mathbb{E}(X) = \sum_{x=1}^{6} x\, Pr(X = x) = 1\frac{1}{6} + 2\frac{1}{6} + 3\frac{1}{6} + 4\frac{1}{6} + 5\frac{1}{6} + 6\frac{1}{6} = \frac{21}{6} = \frac{7}{2} = 3.5\ .$$

o Per la v.c. continua X di cui all'esempio 9.6, il valore medio è:

$$\mathbb{E}(X) = \int_0^1 x\, f(x)\, dx = \int_0^1 x\,(2x)\, dx = \int_0^1 2x^2\, dx = \frac{2}{3}.$$

□

Si può anche definire il valore medio di qualsiasi funzione (misurabile) della v.c. X, e, in particolare, si possono definire i valori medi di $X^2, X^3, \ldots, X^r, \ldots$ In tal caso, i valori medi:

$$\mathbb{E}(X),\ \mathbb{E}(X^2),\ \mathbb{E}(X^3),\ \ldots,\ \mathbb{E}(X^r),\ \ldots$$

si chiamano *momenti della v.c.* X, e cioè momento primo, secondo, terzo, ..., r-esimo, ..., rispettivamente.

o Tra le numerose proprietà dei momenti, ricordiamo il seguente importante teorema:

Se esiste (finito) il valore medio $\mathbb{E}(X^r)$, *allora esistono anche i valori medi* $\mathbb{E}(X^s)$, *per tutti i valori di* $s \leq r$.

Per semplicità di notazione, il valore medio $\mathbb{E}(X)$ viene indicato anche con il simbolo μ; tuttavia, *il valore medio non deve essere confuso con la media aritmetica* (introdotta nel paragrafo 4.3). Infatti, la media aritmetica costituisce una sintesi della posizione di una distribuzione di frequenza, mentre il valore medio riguarda la sintesi di una v.c. e concerne l'aspettativa che si ha su di essa in termini di equivalente certo della distribuzione di probabilità.

o La terminologia inglese per valore medio è «*expectation, mean value, mean*» mentre si parla di «*average*» per riferirsi alla media aritmetica. Quindi, rispetto alla lingua italiana, la distinzione tra le due quantità riesce più agevole.

Si può dimostrare che $\mathbb{E}(.)$ è un *operatore lineare*, cioè per ogni α e β costanti reali, si ha:

$$\mathbb{E}(\alpha X + \beta Y) = \alpha\,\mathbb{E}(X) + \beta\,\mathbb{E}(Y),$$

se esistono i valori medi di X e di Y.

Ponendo $\beta = 0$, si ha:

$$\mathbb{E}(\alpha X) = \alpha\,\mathbb{E}(X).$$

Invece, se si pone $\alpha = \beta = 1$, oppure $\alpha = 1,\ \beta = -1$, si ha:

$$\mathbb{E}(X \pm Y) = \mathbb{E}(X) \pm \mathbb{E}(Y),$$

che mostra come *il valore medio di una somma (differenza) è uguale alla somma (differenza) dei valori medi*, se questi esistono.

Poiché una costante c è null'altro che una v.c. discreta X che assume solo il valore c con probabilità uguale a 1, per calcolare il valore medio di una costante basta osservare che:

$$\mathbb{E}(c) = c\,P_r\,(X = c) = c\,1 = c$$

per cui *il valore medio di una costante è proprio la costante*: $\mathbb{E}(c) = c$.

La relazione di linearità per l'operatore $\mathbb{E}(.)$ si può generalizzare ad una successione di costanti α_i e di v.c. X_i tali che esistano i rispettivi valori medi:

$$\mathbb{E}\left(\sum_{i=1}^{\infty} \alpha_i X_i\right) = \sum_{i=1}^{\infty} \alpha_i\,\mathbb{E}(X_i).$$

o È importante osservare che le proprietà del valore medio permangono valide sia per v.c. discrete che continue, anche se il valore medio è espresso mediante una somma nel primo caso ed un integrale definito nel secondo caso; quest'unica formulazione semplifica molti sviluppi successivi.

Se alla v.c. X si sottrae la costante $\mu = \mathbb{E}(X)$, la v.c. $X - \mu$ viene definita *v.c. scarto* (oppure, v.c. scarto dal valore medio).

È agevole dimostrare che il valore medio della v.c. scarto è sempre 0. Infatti,

$$\mathbb{E}(X - \mu) = \mathbb{E}(X) - \mathbb{E}(\mu) = \mu - \mu = 0.$$

D'altra parte, se la v.c. scarto $X - \mu$ assume valori molto piccoli con grande probabilità, allora la distribuzione di X sarà tutta accentrata attorno al suo valore medio (e quindi l'aspettativa $\mathbb{E}(X)$ appare molto affidabile nel sintetizzare la v.c. X). Per contro, se la v.c. $X - \mu$ assume valori molto elevati (e quindi distanti da μ) con elevate probabilità, allora la distribuzione di X sarà piuttosto dispersa rispetto al suo valore medio (e quindi si verificheranno spesso valori molto distanti da $\mathbb{E}(X)$). Dovendo scommettere sui possibili valori di X, nel primo caso i valori attorno al valore medio saranno «affidabili» mentre nel secondo caso aumenta il rischio se ci si concentra attorno al valore medio.

In breve, tale ragionamento mostra che la distribuzione della v.c. $X - \mu$ consente di valutare il *rischio* (in termini di credibilità, affidabilità, variabilità, dispersione, etc.) connesso alla v.c. X. Poiché il valore medio di $X - \mu$ è sempre 0, occorre sintetizzare il rischio della v.c. X mediante il valore medio di $(X - \mu)^2$. Infatti, tale quantità è 0 se e solo se la v.c. assume il valore μ con probabilità 1, mentre assume un valore sempre più grande se la v.c. X si disperde sempre più rispetto al suo valore medio (cioè alla sua aspettativa).

Pertanto, si definisce ***varianza*** della v.c. X la quantità:

$$Var(X) = \mathbb{E}(X-\mu)^2 = \begin{cases} \sum_{i=1}^{\infty} (x_i - \mu)^2\, p_i & \text{se } X \text{ è una v.c. discreta;} \\ \int_{-\infty}^{+\infty} (x-\mu)^2\, f(x)dx, & \text{se } X \text{ è una v.c. continua.} \end{cases}$$

Tale quantità è sempre ben definita se esiste finito il valore medio di X^2.

o Purtroppo, anche la quantità $Var(X)$ viene spesso indicata con lo stesso simbolo σ^2 con il quale si indica la varianza di una distribuzione di frequenza. Tuttavia, come già

esplicitato per il valore medio, le due quantità sono concettualmente ben diverse perché *la varianza di una v.c. è una misura probabilistica del rischio connesso alla prova.*

o In analogia a quanto definito nel paragrafo 5.2, la radice quadrata della varianza si chiama *scarto quadratico medio*, e viene indicata con σ (oppure, con $\sigma(X)$).

La varianza di una v.c. gode di alcune proprietà, tutte immediatamente derivabili dalla sua definizione, e che brevemente riassumiamo:

i) $Var(X) = 0$ se e solo se $Pr(X = c) = 1$. In tal caso, la v.c. X si chiama *degenere* perché assume un solo valore (c) con probabilità 1. Ovviamente, la quantità c è anche il valore medio di tale v.c. degenere.

ii) $Var(cX) = c^2\, Var(X)$. Quindi, moltiplicando una v.c. per una costante la sua varianza viene moltiplicata per il quadrato della costante.

iii) $Var(X \pm c) = Var(X)$. Quindi, l'aggiunta o la sottrazione di una costante non modifica la varianza di una v.c.

iv) $Var(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$. Quindi, la varianza di una v.c. si può calcolare sottraendo al valore medio del quadrato della v.c. il quadrato del suo valore medio, ovvero *la varianza è uguale al momento secondo meno il momento primo al quadrato.*

Grazie al teorema di esistenza dei momenti (prima enunciato), se esiste (finito) $\mathbb{E}(X^2)$ allora anche $\mathbb{E}(X)$ sarà finito e, alla fine, esisterà finita anche la varianza, sulla base dell'ultima proprietà.

Inoltre, l'ultima proprietà è particolarmente utile nel calcolo della varianza perché evita di specificare preliminarmente la v.c. scarto ed i valori che essa assume, come vedremo nell'esempio seguente.

Esempio 9.9

o Per la v.c. discreta X di cui all'esempio 9.5, abbiamo già calcolato il valore medio (pari a $\mu = \frac{7}{2}$) per cui la varianza risulta così definita:

$$\begin{aligned} Var(X) = \mathbb{E}\left(X - \frac{7}{2}\right)^2 &= \sum_{x=1}^{6} \left(x - \frac{7}{2}\right)^2 Pr(X = x) \\ &= (1 - \tfrac{7}{2})^2 \tfrac{1}{6} + (2 - \tfrac{7}{2})^2 \tfrac{1}{6} + \cdots + (6 - \tfrac{7}{2})^2 \tfrac{1}{6} = \frac{35}{12}. \end{aligned}$$

Tuttavia, ricordando la proprietà iv) della varianza, basta calcolare il valore medio di X^2, cioè:

$$\mathbb{E}(X^2) = \sum_{x=1}^{6} (x^2)\, Pr(X = x) = 1^2 \tfrac{1}{6} + 2^2 \tfrac{1}{6} + \cdots + 6^2 \tfrac{1}{6} = \frac{91}{6} = 15.167$$

e quindi:

$$Var(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12} = 2.917\ .$$

o Procedendo in modo analogo, per calcolare la varianza della v.c. continua X di cui all'esempio 9.6 (il cui valore medio $\mu = 2/3$ è stato già ottenuto nell'esempio 9.8), basta ricavare anzitutto il momento secondo (cioè il valore medio della v.c. X^2):

$$\mathbb{E}(X^2) = \int_0^1 x^2 f(x)\, dx = \int_0^1 x^2 (2x)\, dx = \int_0^1 2x^3\, dx = \frac{1}{2}.$$

e quindi:

$$Var(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{18}.$$

□

Se la v.c. X non è degenere e possiede valor medio $\mathbb{E}(X) = \mu$ e varianza $Var(X) = \sigma^2 < +\infty$, si può definire la ***v.c. standardizzata*** Z mediante:

$$Z = \frac{X - \mathbb{E}(X)}{\sqrt{Var(X)}} = \frac{X - \mu}{\sigma}.$$

È agevole dimostrare che:

$$\mathbb{E}(Z) = 0; \qquad Var(Z) = 1.$$

Inoltre, il *coefficiente di asimmetria* e il *coefficiente di curtosi* della v.c. X sono pari ai valori medi di Z^3 e Z^4, rispettivamente:

$$Asym(X) = \mathbb{E}(Z^3) = \mathbb{E}\left(\frac{X - \mu}{\sigma}\right)^3; \quad Kurt(X) = \mathbb{E}(Z^4) = \mathbb{E}\left(\frac{X - \mu}{\sigma}\right)^4;$$

con significato e interpretazione analoghi alle misure introdotte, in ambito descrittivo, nel capitolo 6.

Data una v.c. X che possiede finito il momento quarto, le quantità seguenti:

$\mathbb{E}(X)$;	$Var(X)$;	$Asym(X)$;	$Kurt(X)$;

sono definite *momenti caratteristici* della v.c. X, perché la loro conoscenza fornisce adeguate informazioni sull'aspettativa, sul rischio e sulla forma dell'intera distribuzione della v.c.

Alcuni teoremi concernenti il valore medio e la varianza di una combinazione lineare di v.c. sono particolarmente importanti per le successive utilizzazioni inferenziali.

In modo intuitivo, osserviamo che una successione di v.c. consiste di ***v.c. indipendenti*** se qualsiasi valore assunto da ciascuna di esse non modifica la distribuzione di probabilità di nessuna delle altre.

- Abbiamo già visto che *il valore medio della combinazione lineare* di una qualsiasi successione $(X_1, X_2, \ldots)$ di v.c. è sempre uguale alla combinazione lineare dei corrispondenti valori medi, cioè:

$$\mathbb{E}(\alpha_1 X_1 + \alpha_2 X_2 + \ldots) = \alpha_1 \mathbb{E}(X_1) + \alpha_2 \mathbb{E}(X_2) + \ldots$$

- Invece, *la varianza di una combinazione lineare* di *v.c. indipendenti* è uguale alla combinazione lineare delle corrispondenti varianze con i quadrati dei coefficienti. Pertanto, la seguente relazione vale *solo se le v.c. sono indipendenti*:

$$Var(\alpha_1 X_1 + \alpha_2 X_2 + \ldots) = (\alpha_1)^2 Var(X_1) + (\alpha_2)^2 Var(X_2) + \ldots$$

In aggiunta ai valori medi, esistono altre misure sintetiche di una v.c. e tra esse ci limitiamo a citare brevemente le seguenti.

- La ***moda*** di una v.c. discreta è il valore Mo (se esiste e se è unico) più probabile; se la v.c. è continua è quel valore Mo tale che la funzione di densità presenta la proprietà: $f(Mo) \geq f(x), \forall x$, ovvero la funzione di densità presenta un unico massimo in $x = Mo$.

- La ***mediana*** di una v.c. è quel valore $Me(X)$ tale che:

$$Pr(X \leq Me(X)) \geq \frac{1}{2}; \quad Pr(X \geq Me(X)) \geq \frac{1}{2}.$$

  Ovviamente, per una v.c. X continua, risulta: $F(Me(X)) = 0.5$.

- Il ***quantile di ordine p*** di una v.c. è quel valore x_p definito da:

$$Pr(X \leq x_p) \geq p; \quad Pr(X \geq x_p) \leq 1 - p, \ \forall p \in (0,1).$$

  Per una v.c. X continua, risulta: $F(x_p) = p$. Quando $p = 0.25; 0.10; 0.01$, i quantili si chiamano *quartili, decili, percentili*, rispettivamente.

- L'*intervallo interquartile* di una v.c. X è definito da: $IQR(X) = x_{0.75} - x_{0.25}$ e fa riferimento ad un intervallo di valori, centrale nella distribuzione di X, che può verificarsi con probabilità 0.5.

Infine, uno dei risultati più importanti del Calcolo delle probabilità è la disuguaglianza di Cebicev, perché è una relazione valida per tutte le v.c. che possiedono finito il momento secondo.

Più precisamente, la ***disuguaglianza di Cebicev*** afferma che, per ogni v.c. X tale che $\mathbb{E}(X) = \mu$, $Var(X) = \sigma^2 < +\infty$, si ha:

$$\boxed{Pr(|X - \mu| < \epsilon, \sigma) \geq 1 - \frac{1}{\epsilon^2}}.$$

Posto $\epsilon, \sigma = k$, la precedente espressione è equivalente alla seguente:

$$\boxed{Pr(|X - \mu| < k) \geq 1 - \frac{\sigma^2}{k^2}}.$$

- In breve, la disuguaglianza di Cebicev afferma che la probabilità che qualsiasi v.c. (discreta o continua) assuma un valore nell'intervallo $\mu \pm \epsilon, \sigma$ è sempre maggiore di $1 - 1/\epsilon^2$. Quindi, più è ampio l'intervallo rispetto a σ più la probabilità tende a 1.
- Essa è valida per ogni $\epsilon > 0$, ma presenta qualche utilità solo se $\epsilon > 1$. Infatti, se $\epsilon = 1$, la disuguaglianza di Cebicev si risolve in: $Pr(|X - \mu| < \sigma) \geq 0$, che equivale al 1° postulato, mentre se $\epsilon < 1$, risulta banale affermare che la probabilità di un evento è non inferiore ad un numero negativo.
- Se si pone $\epsilon = 2$ nella disuguaglianza di Cebicev si ha: $Pr(|X - \mu| < 2\sigma) \geq 0.75$, cioè per qualsiasi v.c. è superiore (o uguale) a 3/4 la probabilità che la v.c. X assuma valori che non si discostano dal valore medio μ più di 2 volte σ.
- Ovviamente, se si conosce la distribuzione di probabilità o la funzione di densità della v.c. X si riesce a calcolare esattamente le probabilità in questione; tuttavia, l'importanza della disuguaglianza di Cebicev risiede nel fatto che essa vale senza alcuna conoscenza aggiuntiva sulla v.c., eccetto quella del valore medio e della varianza. Nello specifico, si è anzi dimostrato che non è possibile migliorare tale disuguaglianza senza aggiungere altre informazioni sulla v.c.

9.7 Variabili casuali multiple

La problematica sinora affrontata è stata ricondotta allo studio di un numero reale associato ad un evento, ma è concettualmente possibile associare ad ogni evento anche una coppia, tripla, ..., m-pla ordinata di numeri reali, e studiarne poi la relativa distribuzione di probabilità.

- Se dal data set STUDENTI si estrae casualmente un soggetto di cui si rileva peso, altezza ed età si sta effettuando una prova che genera una terna ordinata di numeri reali. Per esempio, l'estrazione dello studente Mario dà luogo alla terna di valori $(173, 78, 19)$, mentre quella della studentessa Maria dà luogo alla terna $(165, 59, 20)$.
- Similmente, se viene estratta casualmente una foglia dal data set CAMELLIA e si è interessati alla lunghezza e larghezza di tale foglia, si genera un evento cui è associata una coppia ordinata di numeri reali (x, y) che misurano, rispettivamente, la lunghezza X e la larghezza Y della foglia prescelta.

Ciò che distingue in modo sostanziale una v.c. multivariata (o v.c. multipla) da una v.c. univariata (o v.c. semplice) è il fatto che non interessa studiare solo la distribuzione di probabilità di ciascuna delle componenti ma, soprattutto, la probabilità del contemporaneo verificarsi delle coppie, triple, ..., m-ple di numeri reali. Ciò avviene per valutare associazioni, connessioni, concordanze e relazioni tra le varie componenti di una v.c. multivariata.

Sia definita una prova $\mathcal{P}$ che genera una collezione di eventi inclusi in uno spazio campione Ω. Se ad ogni evento E si associa una unica m-pla ordinata di numeri reali $(x_1, x_2, \ldots, x_m)$ si definisce una ***v.c. multivariata*** $(X_1, X_2, \ldots, X_m)$ ad m dimensioni. In pratica, ad ogni evento $E \subset \Omega$ corrisponde un unico punto di

$\mathbb{R}^m$. Perché tale corrispondenza sia lecita, è necessario che la relazione tra spazio campione Ω e spazio reale $\mathbb{R}^m$ sia una *funzione misurabile.*

Le v.c. $X_1, X_2, \ldots, X_m$ si dicono ***v.c. componenti*** della v.c. multivariata. Ogni sotto-insieme di $k < m$ v.c. componenti, derivato da una v.c. multivariata di dimensione m, si chiama ***v.c. marginale*** di dimensione $k = 1, 2, \ldots$ Ogni v.c. componente è, evidentemente, una v.c. marginale di dimensione $k = 1$.

In corrispondenza della natura dello spazio campione, avremo *v.c. multivariate discrete o continue* se, rispettivamente, tutte le v.c. componenti sono v.c. discrete o continue.

Lo sviluppo coerente dei problemi che le v.c. multivariate pongono imporrebbe un impegno notevole soprattutto sul piano della formalizzazione, non sempre giustificato dal livello cui si pone questo volume. Per questo, rinviando alla bibliografia per ogni necessario approfondimento, ci limitiamo nel paragrafo seguente a sviluppare con un certo dettaglio il solo caso di una v.c. doppia discreta; resta inteso che tale discussione contiene, in ogni caso, la gran parte delle problematiche presenti in ogni ulteriore generalizzazione.

Come per il caso univariato, nell'assegnazione di una probabilità, occorre distinguere il caso di una distribuzione di probabilità multivariata per v.c. discrete e quello di una funzione di densità multivariata per v.c. continue.

In entrambi i casi, occorre rispondere alle seguenti domande:

i) quando una v.c. multivariata è nota?

ii) quando una v.c. multivariata è ben definita?

iii) come si rappresenta una v.c. multivariata?

9.8 Variabili casuali doppie

Le *v.c. doppie discrete* sono definite come una corrispondenza tra gli eventi e una coppia ordinata di numeri reali: quindi, una v.c. doppia discreta (X, Y) assume, al più, un numero finito o infinito numerabile di possibili coppie ordinate di valori reali. L'insieme di valori (x_i, y_j) tali che la probabilità del loro verificarsi è positiva si chiama ***supporto*** della v.c. doppia (X, Y).

Una v.c. doppia discreta (X, Y) è nota quando, in corrispondenza di tutte le possibili coppie ordinate $(x_i, y_j), i = 1, 2, \ldots; j = 1, 2, \ldots$, sono note le probabilità p_{ij} del verificarsi del valore x_i per X e *contemporaneamente* del valore y_j per Y. In altri termini, una v.c. doppia discreta è nota quando sono note tutte le probabilità:

$$p_{ij} = Pr\left((X = x_i) \cap (Y = y_j)\right), \; i = 1, 2, \ldots; j = 1, 2, \ldots$$

L'insieme dei valori (x_i, y_j) e delle corrispondenti probabilità p_{ij} si chiama ***distribuzione di probabilità della v.c. doppia discreta*** (X, Y).

Una v.c. doppia discreta (X, Y) è ben definita se e solo se la distribuzione di probabilità soddisfa le due seguenti condizioni:

$$p_{ij} \geq 0, \; \forall (i, j) \in \{i = 1, 2, \ldots; j = 1, 2, \ldots\}; \qquad \sum_i \sum_j p_{ij} = 1 .$$

Ipotizzando che la *v.c. componente* X possa assumere i valori reali $x_1, x_2, \ldots, x_k$ e che la *v.c. componente* Y possa assumere i valori reali $y_1, y_2, \ldots, y_h$, allora è comune (ed è anche conveniente) presentare una distribuzione di probabilità per la v.c. doppia discreta (X, Y) mediante una tabella del tipo seguente:

$X \backslash\backslash Y$	y_1	y_2	$\ldots$	y_j	$\ldots$	y_h	Totali
x_1	p_{11}	p_{12}	$\ldots$	p_{1j}	$\ldots$	p_{1h}	$p_{1.}$
x_2	p_{21}	p_{22}	$\ldots$	p_{2j}	$\ldots$	p_{2h}	$p_{2.}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_i	p_{i1}	p_{i2}	$\ldots$	p_{ij}	$\ldots$	p_{ih}	$p_{i.}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_k	p_{k1}	p_{k2}	$\ldots$	p_{kj}	$\ldots$	p_{kh}	$p_{k.}$
Totali	$p_{.1}$	$p_{.2}$	$\ldots$	$p_{.j}$	$\ldots$	$p_{.h}$	1

Le probabilità poste ai due margini della tabella doppia sono ottenute sommando le probabilità di ciascuna riga o colonna, rispettivamente. Esse sono dette ***probabilità marginali*** ed esplicitano la distribuzione di probabilità delle due v.c. componenti X e Y della v.c. doppia (X, Y).

- Dal punto di vista mnemonico, si utilizzano le stesse regole già discusse nel caso di una distribuzione doppia di variabili statistiche (paragrafo 7.2). Per esempio, $p_{2.}$ è la probabilità dell'evento $(X = x_2)$, definita per la v.c. X senza tenere conto dei valori assunti dalla v.c. Y; similmente, $p_{.1}$ è la probabilità dell'evento $(Y = y_1)$, definito per la v.c. Y senza tenere conto dei valori assunti dalla v.c. X.
- Data una v.c. discreta doppia (X, Y), si può ricavare la distribuzione (univariata) della v.c. marginale X (cioè, la prima componente) sommando la distribuzione di probabilità doppia rispetto ai valori che la v.c. Y assume. Similmente, si può ricavare la distribuzione (univariata) della v.c. marginale Y (cioè, la seconda componente) sommando la distribuzione di probabilità doppia rispetto ai valori che la v.c. X assume.
- Sul piano formale, le probabilità per v.c. doppie e le probabilità delle v.c. componenti sono connesse dalle seguenti relazioni:

$$p_{i.} = \sum_{j=1}^{h} p_{ij}, \;\; \forall i = 1, 2, \ldots, k; \qquad p_{.j} = \sum_{i=1}^{k} p_{ij}, \;\; \forall j = 1, 2, \ldots, h.$$

- La rappresentazione grafica di una distribuzione di probabilità è possibile in uno spazio a 3 dimensioni, ponendo sulle prime due coordinate i valori (x_i, y_j) assunti, rispettivamente, dalle due v.c. componenti (X, Y) e, sulla terza coordinata, le probabilità corrispondenti p_{ij}. Nell'esempio 9.10 presenteremo una visualizzazione di tale rappresentazione grafica.

Data una v.c. doppia discreta (X, Y), è possibile studiare la distribuzione di probabilità di una delle due v.c. componenti dopo aver fissato il valore assunto dall'altra v.c. componente.

Da una v.c. doppia si possono dedurre due gruppi di v.c. condizionate, cioè:

i) la v.c. $Y|(X = x_0)$ nella quale la v.c. componente Y è condizionata al valore x_0 prefissato ed appartenente al supporto della v.c. X;

ii) la v.c. $X|(Y = y_0)$ nella quale la v.c. componente X è condizionata al valore y_0 prefissato ed appartenente al supporto della v.c. Y.

Ricordando la definizione di probabilità condizionata (paragrafo 8.11), per un certo fissato valore di x_0, tale che $Pr(X = x_0) > 0$, avremo che la ***distribuzione di probabilità della v.c. condizionata*** $Y|(X = x_0)$ è definita da:

$$Pr(Y = y_j|X = x_0) = \frac{Pr((Y = y_j) \cap (X = x_0))}{Pr(X = x_0)}, \quad \forall j = 1, 2, \ldots, h.$$

Similmente, per un certo fissato valore di y_0, tale che $Pr(Y = y_0) > 0$, avremo che la ***distribuzione di probabilità della v.c. condizionata*** $X|(Y = y_0)$ è definita da:

$$Pr(X = x_i|Y = y_0) = \frac{Pr((X = x_i) \cap (Y = y_0))}{Pr(Y = y_0)}, \quad \forall i = 1, 2, \ldots, k.$$

o Si noti che la *distribuzione condizionata* è quella di una v.c. univariata ottenuta da una v.c. doppia imponendo ad una delle due v.c. componenti di assumere un valore prefissato. Invece, una *distribuzione marginale* è anch'essa una v.c. univariata ottenuta dalla v.c. doppia esaminando solo i valori dell'altra componente.

o Naturalmente, essendo delle v.c. univariate ben definite, sia per le v.c. componenti che per le v.c. condizionate è possibile il calcolo del valore medio, della varianza e di qualsiasi altro indicatore sintetico (mediana, moda, quartili, etc.). In quest'ultimo caso, si parla di valore medio condizionato, varianza condizionata, e così via.

Sul piano formale, le definizioni sono immediate; per esempio, *il valore medio della v.c. Y condizionato al valore* $(X = x_0)$ è definito da:

$$\mathbb{E}(Y|X = x_0) = \sum_{j=1}^{h} y_j \, Pr(Y = y_j|X = x_0).$$

L'esempio seguente illustra la definizione, il calcolo e le interpretazioni di una v.c. doppia discreta e delle v.c. marginali e condizionate che da essa si possono derivare, proponendo anche un'adeguata rappresentazione grafica.

Esempio 9.10

La prova consiste nell'estrarre una pallina da ciascuna di 3 urne, ognuna delle quali contiene 3 palline omogenee numerate con i numeri 1, 2, 3, e di osservare i tre numeri risultanti. Su tale prova definiamo una v.c. doppia (X, Y) nel modo seguente:

X = *la somma dei numeri sulle tre palline estratte*;
Y = *la differenza tra il massimo e il minimo numero fra i tre estratti.*

Per studiare correttamente tale v.c. esaminiamo tutti i $3^3 = 27$ possibili risultati di questa prova, ove per semplicità di notazione indichiamo – per esempio – con (1,2,2) l'evento che consiste nell'estrarre dalle tre urne tre palline con i numeri 1, 2 e 2, ordinatamente, e così via:

(1,1,1)	(1,1,2)	(1,1,3)	(1,2,1)	(1,2,2)	(1,2,3)	(1,3,1)	(1,3,2)	(1,3,3)
(2,1,1)	(2,1,2)	(2,1,3)	(2,2,1)	(2,2,2)	(2,2,3)	(2,3,1)	(2,3,2)	(2,3,3)
(3,1,1)	(3,1,2)	(3,1,3)	(3,2,1)	(3,2,2)	(3,2,3)	(3,3,1)	(3,3,2)	(3,3,3)

Trattandosi di una partizione equiprobabile dello spazio campione, ciascuno di tali eventi possiede probabilità pari a $\frac{1}{27}$.

È agevole indicare, per ciascuno di tali eventi, la coppia ordinata di valori (x, y) di numeri reali determinati dalla v.c. doppia (X, Y). Per esempio, all'evento (1,3,1) corrisponde la coppia (5,2) perché la somma X dei tre numeri estratti assume il valore $x = 1+3+1 = 5$ mentre la differenza Y tra il massimo e il minimo tra i tre numeri estratti assume il valore $y = 3 - 1 = 2$. In tal modo, nello stesso ordine degli eventi sopra riportati, la v.c. (X, Y) assume come valori le seguenti coppie ordinate:

(3,0)	(4,1)	(5,2)	(4,1)	(5,1)	(6,2)	(5,2)	(6,2)	(7,2)
(4,1)	(5,1)	(6,2)	(5,1)	(6,0)	(7,1)	(6,2)	(7,1)	(8,1)
(5,2)	(6,2)	(7,2)	(6,2)	(7,1)	(8,1)	(7,2)	(8,1)	(9,0)

A ciascuna di tali coppie, evidentemente, è assegnata probabilità $\frac{1}{27}$ di verificarsi.

Ora, per studiare la distribuzione della v.c. (X, Y) occorre tenere presente che X assume i valori $x = 3, 4, \ldots, 9$ mentre Y assume i valori $y = 0, 1, 2$. Pertanto, la distribuzione doppia della v.c. (X, Y), agevolmente derivabile dal riquadro precedente, è quella della tabella seguente.

$X \backslash\backslash Y$	$Y = 0$	$Y = 1$	$Y = 2$	Totali
$X = 3$	$\frac{1}{27}$	0	0	$\frac{1}{27}$
$X = 4$	0	$\frac{3}{27}$	0	$\frac{3}{27}$
$X = 5$	0	$\frac{3}{27}$	$\frac{3}{27}$	$\frac{6}{27}$
$X = 6$	$\frac{1}{27}$	0	$\frac{6}{27}$	$\frac{7}{27}$
$X = 7$	0	$\frac{3}{27}$	$\frac{3}{27}$	$\frac{6}{27}$
$X = 8$	0	$\frac{3}{27}$	0	$\frac{3}{27}$
$X = 9$	$\frac{1}{27}$	0	0	$\frac{1}{27}$
Totali	$\frac{3}{27}$	$\frac{12}{27}$	$\frac{12}{27}$	1

L'interpretazione di tale tabella è immediata. Per esempio,

$$Pr((X = 4),\ (Y = 1)) = Pr((X = 4) \cap (Y = 1)) = \frac{3}{27}.$$

perché la somma assume il valore 4 e *contemporaneamente* la differenza tra massimo e minimo assume il valore 1 se e solo se si verificano gli eventi:

$$(1, 1, 2) \quad \text{oppure} \quad (1, 2, 1) \quad \text{oppure} \quad (2, 1, 1),$$

cioè 3 eventi elementari di una partizione equiprobabile di 27 eventi elementari.

Similmente, risulta che $[(X = 8) \cap (Y = 2)]$ è un evento impossibile; e così via, per tutti gli altri.

La precedente tabella mostra la distribuzione di probabilità della v.c. doppia (X, Y) ed esaurisce ogni conoscenza riguardante tale v.c. Grazie a tale tabella, possiamo dedurre qualsiasi affermazione probabilistica riguardante la prova in esame.

La v.c. è ben definita, perché tutte le probabilità sono non-negative e la loro somma è 1; inoltre, la sua rappresentazione grafica (Figura 9.6) consiste in aste di altezze proporzionali

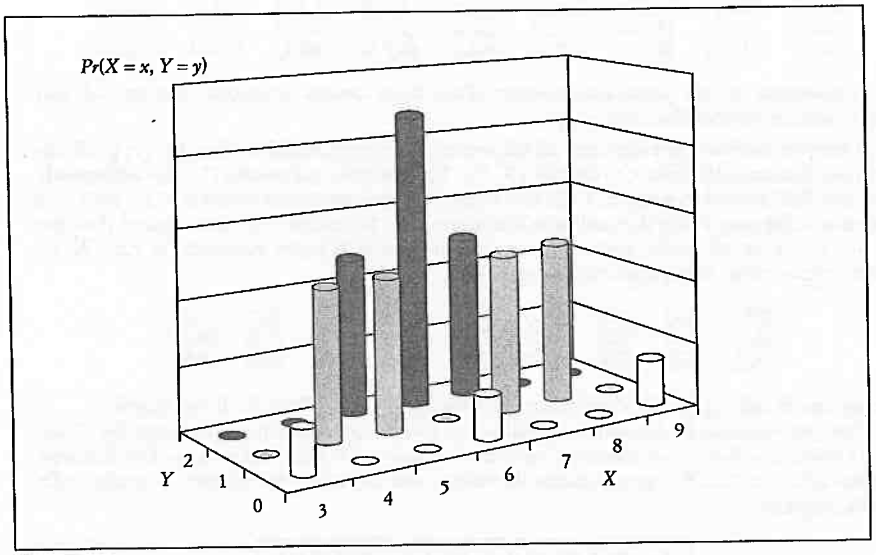

FIGURA 9.6. Distribuzione di probabilit di una variabile casuale doppia discreta.

alle probabilità definite sui soli 9 punti del piano (X,Y) con probabilità non nulla, cioè sul *supporto* di (X,Y).

Dalla conoscenza della distribuzione doppia si deducono agevolmente le distribuzioni delle v.c. marginali X oppure Y, sommando ciascuna riga cioè rispetto alle colonne (rispetto ai valori di Y) oppure sommando ciascuna colonna cioè rispetto alle righe (rispetto ai valori di X).

Le *v.c. componenti* X e Y di una v.c. doppia (X,Y) (dette anche *v.c. marginali* perché le rispettive distribuzioni sono ai margini della tabella doppia di probabilità) sono anch'esse necessariamente ben definite e costituiscono, ciascuna di esse, una v.c. univariata. Per esempio, dalla precedente tabella si vede che la v.c. Y assume i valori $(Y=0), (Y=1)$ e $(Y=2)$ con probabilità $\frac{3}{27}, \frac{12}{27}, \frac{12}{27}$, rispettivamente.

Le *v.c. condizionate*, invece, si ottengono quando si fissa il valore assunto da una delle due v.c. componenti e si studia la distribuzione di probabilità dell'altra v.c., dato quel valore. Per esempio, nella prova in esame, potremmo conoscere che la differenza tra il massimo e il minimo è 1 e voler scommettere sulla somma. Evidentemente, l'informazione in nostro possesso – cioè il fatto che l'evento $(Y=1)$ si è verificato – modifica la probabilità dei possibili valori per X. Infatti, senza tale informazione, noi avremmo scommesso sull'evento $(X=6)$, cui corrisponde la probabilità più elevata, ma se si è verificato $(Y=1)$ allora tale valore è impossibile per X.

Ricordando la definizione di probabilità condizionata, si possono derivare tutte le possibili distribuzioni condizionate di interesse; per esempio, dato $y=1$, al variare di $X=3,4,\ldots,9$, si ottiene la distribuzione condizionata della v.c. $(X|Y=1)$. Così,

$$Pr(X=3|Y=1)=\frac{Pr(X=3,\,Y=1)}{Pr(Y=1)}=\frac{0}{\frac{12}{27}}=0;$$

$$Pr(X=5|Y=1)=\frac{Pr(X=5,\,Y=1)}{Pr(Y=1)}=\frac{\frac{3}{27}}{\frac{12}{27}}=\frac{1}{4};$$

e così via, per tutti gli altri valori.

Nella tabella seguente presentiamo la distribuzione di probabilità delle 3 possibili v.c. condizionate $(X|Y=y)$ quando, a turno, fissiamo $y=0$, $y=1$ oppure $y=2$, rispettivamente.

$(X\|Y=y)$	$Pr(X\|Y=0)$	$Pr(X\|Y=1)$	$Pr(X\|Y=2)$
$X=3$	$\frac{1}{3}$	0	0
$X=4$	0	$\frac{1}{4}$	0
$X=5$	0	$\frac{1}{4}$	$\frac{1}{4}$
$X=6$	$\frac{1}{3}$	0	$\frac{1}{2}$
$X=7$	0	$\frac{1}{4}$	$\frac{1}{4}$
$X=8$	0	$\frac{1}{4}$	0
$X=9$	$\frac{1}{3}$	0	0
Totali	1	1	1

L'interpretazione che se ne trae è la seguente: se, per esempio, si è verificato l'evento $(Y=0)$ ciò significa che si è verificato solo uno dei 3 eventi elementari (1,1,1) oppure (2,2,2) oppure (3,3,3); quindi, i soli valori di X con probabilità non nulle sono, rispettivamente, $(X=3)$ oppure $(X=6)$ oppure $(X=9)$ che, essendo equiprobabili, possiedono probabilità $\frac{1}{3}$, come correttamente risulta dalla distribuzione di probabilità della v.c. $(X|Y=0)$.

Si osservi come la v.c. $(X|Y=y)$ sia una v.c. univariata la quale, essendosi Y già verificato, assume tutti i valori di X con probabilità che (generalmente) cambiano perché Y si è verificato. La v.c. condizionata $(X|Y=y)$ è definita sullo stesso supporto della v.c. marginale X, solo che la distribuzione di probabilità generalmente si modifica perché si è verificato un valore per Y.

Quindi, $(X|Y=y)$ è una *v.c. univariata* ben definita, e per essa si possono calcolare il valor medio e la varianza per ciascun prefissato y. Se si nota che le precedenti distribuzioni condizionate sono simmetriche rispetto al valore 6, si deduce subito che:

$$\mathbb{E}(X|Y=0)=\mathbb{E}(X|Y=1)=\mathbb{E}(X|Y=2)=6;$$

come è confermato da un calcolo esplicito di tali valori medi. Per esempio, si ha:

$$\mathbb{E}(X|Y=0)=3\,\frac{1}{3}+4\,(0)+\cdots+9\,\frac{1}{3}=6$$

In modo analogo, si ottiene:

$$\mathbb{E}(X^2|Y=0)=42;\ \ \mathbb{E}(X^2|Y=1)=38.5;\ \ \mathbb{E}(X^2|Y=2)=36.5\ .$$

Ricordando che la varianza di una v.c. X è il valore medio della v.c. X^2 meno il quadrato del valore medio di X, si ha:

$$Var(X|Y=0)=42-6^2=6;\ \ \ Var(X|Y=1)=2.5;\ Var(X|Y=2)=0.5\ .$$

Poiché per la v.c. marginale X risulta $\mathbb{E}(X)=6$; $Var(X)=2$, la considerazione conclusiva è la seguente: il valore medio di X non cambia con la conoscenza del valore

di Y ma cambia la distribuzione di probabilità di X e la sua variabilità. In particolare, la varianza della distribuzione di X, quando $Y = 1$ oppure quando $Y = 2$, diviene notevolmente più bassa, ovvero il rischio connesso ad una scommessa sull'aspettativa diventa molto inferiore.

□

Per gli sviluppi successivi, è fondamentale il concetto di ***indipendenza probabilistica*** tra v.c., che si ricollega immediatamente alla definizione di indipendenza tra eventi (paragrafo 8.11) e che ora specifichiamo per v.c. doppie discrete.

Una v.c. doppia (X, Y) ha *componenti indipendenti* se e solo se, per qualsiasi coppia di valori reali x_i, y_j si ha:

$$Pr((X = x_i) \cap (Y = y_j)) = Pr(X = x_i)\ Pr(Y = y_j)\ .$$

Quindi, due v.c. X e Y sono ***indipendenti*** se la distribuzione doppia di probabilità si fattorizza identicamente nel prodotto delle distribuzioni di probabilità marginali.

È agevole dedurre che la precedente definizione di indipendenza implica che la distribuzione di probabilità condizionata di una v.c. componente coincide identicamente con la distribuzione della v.c. componente, ovvero – in altri termini – il valore condizionante di una v.c. non modifica la probabilità dell'altra v.c.

Quando, invece, le componenti non sono indipendenti è interessante cercare di sintetizzare la distribuzione doppia di probabilità tramite valori medi che tengano conto del *contemporaneo* verificarsi dei valori di entrambe le componenti della v.c. doppia. In tale direzione, *il valore medio del prodotto* XY delle v.c. componenti della v.c. doppia discreta (X, Y) è definito da:

$$\mathbb{E}(XY) = \sum_{i=1}^{k} \sum_{j=1}^{h} x_i\, y_j\, p_{ij}\ .$$

Esso è ben definito se esistono finite le varianze di entrambe le componenti. Si può dimostrare che *se X e Y sono v.c. indipendenti, allora il valore medio del prodotto è il prodotto dei singoli valori medi.* Formalmente,

$$\mathbb{E}(XY) = \mathbb{E}(X)\,\mathbb{E}(Y).$$

Va sottolineato che non vale il viceversa: cioè, può avvenire che il valore medio del prodotto sia pari al prodotto dei valori medi senza che ciò implichi necessariamente la indipendenza tra le due v.c.

Esempio 9.11

Calcoliamo il valore medio del prodotto delle v.c. componenti per la v.c. doppia discreta esaminata nell'esempio 9.10. A tal fine, applichiamo la definizione di $\mathbb{E}(XY)$ alla distribuzione di probabilità presentata nella prima tabella del precedente esempio; si ottiene:

$$\mathbb{E}(XY) = (3)(0)\frac{1}{27} + (3)(1)0 + (3)(2)0 + (4)(0)0 + (4)(1)\frac{3}{27} + \cdots + (9)(2)0 = \frac{216}{27} = 8\,.$$

Poiché $\mathbb{E}(X) = 6$, $\mathbb{E}(Y) = \frac{4}{3}$, per questa v.c. doppia, il valore medio del prodotto delle v.c. X e Y coincide con il prodotto dei valori medi delle v.c. componenti ma, come è evidente, le due componenti X e Y *non* sono v.c. *indipendenti.*

□

La discussione precedente e l'esempio di una v.c. doppia discreta ha mostrato che dalla conoscenza della v.c. doppia (così come emerge dalla distribuzione doppia di probabilità) si può sempre derivare quella di entrambe le v.c. componenti.

Invece, il problema inverso, cioè la possibilità di risalire dalla conoscenza delle v.c. componenti alla conoscenza della v.c. doppia, non è generalmente risolvibile, tranne quando si tratta di v.c. indipendenti perché, allora, la distribuzione doppia è semplicemente ottenuta (secondo la definizione di indipendenza) mediante il prodotto delle probabilità univariate (cioè marginali) delle due v.c. componenti.

L'esempio seguente mostra la impossibilità di risalire in modo univoco dalla conoscenza delle distribuzioni delle v.c. componenti a quella doppia.

Esempio 9.12

Definiamo una v.c. doppia discreta (X, Y) mediante la seguente distribuzione doppia di probabilità:

$X \backslash\backslash Y$	$Y = y_1$	$Y = y_2$	Totali
$X = x_1$	$\frac{1}{4} - \alpha$	$\frac{1}{4} + \alpha$	$\frac{1}{2}$
$X = x_2$	$\frac{1}{4} + \alpha$	$\frac{1}{4} - \alpha$	$\frac{1}{2}$
Totali	$\frac{1}{2}$	$\frac{1}{2}$	1

È agevole convincersi che, per qualsiasi numero reale $\alpha \in [0, 1/4]$, la precedente distribuzione doppia è ben definita. Quindi, esistono infinite v.c. doppie (una per ogni valore fissato di α) che posseggono le stesse v.c. componenti; quindi, la conoscenza delle v.c. marginali non consente di risalire univocamente alla v.c. doppia, perché da esse (come si vede) non è possibile dedurre alcuna informazione sul valore di α, che è invece necessario per specificare la v.c. doppia.

Osserviamo che, in questo specifico esempio, se fosse $\alpha = 0$ allora le v.c. X e Y sarebbero indipendenti.

□

9.9 Covarianza e correlazione tra variabili casuali

Da un punto di vista più operativo, occorre chiedersi se ed in che misura sia possibile misurare l'eventuale legame probabilistico esistente tra le v.c. componenti di una v.c. doppia discreta.

Una misura delle variazioni congiunte in termini probabilistici tra X e Y può dedursi dal valor medio del prodotto delle v.c. scarto $(X - \mathbb{E}(X))$ e $(Y - \mathbb{E}(Y))$: infatti, se valori concordi in segno hanno maggiore probabilità di valori discordi, tale valore medio sarà positivo; viceversa, il valore medio del prodotto delle v.c.

scarto sarà negativo se vi è una predominanza probabilistica di valori discordi in segno.

Formalmente, data una v.c. doppia discreta (X,Y), i cui valori medi indichiamo con $\mu_x = \mathbb{E}(X)$ e $\mu_y = \mathbb{E}(Y)$ e le cui v.c. componenti possiedono finite le varianze, si definisce ***covarianza*** tra le v.c. X e Y, e si indica con $Cov(X,Y)$, il valore medio del prodotto delle v.c. scarto delle due componenti, ovvero:

$$Cov(X,Y) = \mathbb{E}\left((X-\mu_x)(Y-\mu_y)\right) = \sum_{i=1}^{k}\sum_{j=1}^{h}(x_i-\mu_x)(y_j-\mu_y)\,p_{ij}\,.$$

Semplici sviluppi algebrici nella definizione mostrano che:

$$Cov(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y)\,,$$

che diventa un'espressione molto più agevole per il calcolo, in quanto non richiede la specificazione preliminare delle v.c. scarto. Peraltro, anche in tal caso, si possono utilizzare tutte le proprietà dell'operatore lineare $\mathbb{E}(.)$ presentate nel paragrafo 9.6.

La covarianza misura, quindi, la *variazione congiunta* tra le v.c. componenti di una v.c. doppia, ed è anche indicata con il simbolo σ_{xy}.

- Per qualsiasi costante a,b,c,d, la covarianza gode delle seguenti proprietà:

1. $Cov(X,Y) = Cov(Y,X)$;
2. $Cov(X,X) = Var(X)$;
3. $Cov(a+bX,Y) = Cov(X,a+bY) = b\,Cov(X,Y)$;
4. $Cov(a+bX,c+dY) = b\,d\,Cov(X,Y)$;
5. se X e Y sono *v.c. indipendenti*, allora $Cov(X,Y) = 0$;
6. $Cov(X,Y) \leq \sqrt{Var(X)\,Var(Y)}$;

Esse mostrano che la covarianza è simmetrica rispetto alle due v.c. e che essa generalizza la nozione di varianza; inoltre, la covarianza non cambia per traslazione ma per cambiamenti di scala.

Poiché le proprietà 3. e 4. mostrano che la covarianza dipende dalle unità di misura di X e di Y, è opportuno considerare la covarianza tra le *v.c. componenti standardizzate* definite da:

$$Z_x = \frac{X-\mu_x}{\sigma_x}; \qquad Z_Y = \frac{Y-\mu_y}{\sigma_y}.$$

La covarianza tra le v.c. componenti standardizzate prende il nome di ***coefficiente di correlazione lineare*** (di Bravais e Pearson), e viene indicato con $Corr(X,Y)$, oppure con ρ_{xy} o semplicemente con ρ.

Per la v.c. doppia discreta (X,Y), esso è formalmente definito da:

$$Corr(X,Y) = \mathbb{E}\left(\frac{X-\mu_x}{\sigma_x}\frac{Y-\mu_y}{\sigma_y}\right) = \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}} = \frac{\sigma_{xy}}{\sigma_x\,\sigma_y}.$$

Tenendo conto della definizione di covarianza, tale coefficiente si può esprimere mediante:

$$Corr(X,Y) = \sum_{i=1}^{k}\sum_{j=1}^{h}\frac{(x_i-\mu_x)}{\sigma_x}\frac{(y_j-\mu_y)}{\sigma_y}\,p_{ij}\,.$$

Infine, tenuto conto dei legami tra covarianza, varianza e valori medi, il coefficiente di correlazione si può anche calcolare mediante l'espressione seguente:

$$Corr(X,Y) = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y)}{\sqrt{\left[\mathbb{E}(X^2) - [\mathbb{E}(X)]^2\right]\left[\mathbb{E}(Y^2) - [\mathbb{E}(Y)]^2\right]}}\,.$$

- Le proprietà del coefficiente di correlazione, esposte nel paragrafo 7.5 a proposito di due variabili statistiche, sembrano essere tutte valide anche per il coefficiente di correlazione lineare tra due v.c. In effetti, il coefficiente di correlazione lineare è simmetrico, vale 1 se le due v.c. coincidono, è sempre compreso tra -1 e $+1$ (come si deriva immediatamente dalla proprietà 6. della covarianza, per cui è un indice normalizzato), non cambia se le due v.c. sono trasformate con una funzione di primo grado a coefficienti concordi, e infine vale 0 se X e Y sono v.c. indipendenti.
- Tuttavia, il significato delle varie affermazioni cambia perché esse riguardano le v.c. componenti di una v.c. doppia e non le variabili statistiche. Per esempio, l'affermazione che $Corr(X,Y) = \pm 1$ se e solo se esiste una relazione lineare perfetta tra X e Y va intesa nel senso che la distribuzione doppia delle v.c. (X,Y) è degenere: cioè, che tutti e solo i valori che la v.c. (X,Y) assume sono quelli connessi ad una relazione lineare (cioè di primo grado): $Y = a + b\,X$. In tal caso, lo spazio dei valori possibili non è a 2 dimensioni ma ad 1 dimensione ed i valori possibili per la v.c. (X,Y) si dispongono con probabilità 1 su una retta.

Quindi, l'interpretazione corretta di tale proprietà, in un contesto probabilistico, è la seguente:

$$Corr(X,Y) = \pm 1 \Longrightarrow Pr\,(Y = a + b\,X) = 1\,.$$

- Se X e Y sono v.c. indipendenti, allora: $\mathbb{E}(XY) = \mathbb{E}(X)\,\mathbb{E}(Y)$ per cui:

$$Cov(X,Y) = Corr(X,Y) = 0\,,$$

ma non è necessariamente vero il viceversa. Un coefficiente di correlazione pari a 0 (in tal caso, si dice che X e Y sono ***v.c. incorrelate***) non implica necessariamente la indipendenza tra le v.c. X e Y.

Ciò è ragionevole perché la correlazione misura *un legame probabilistico di tipo lineare* e quindi la incorrelazione (cioè $Corr(X,Y) = 0$) esclude solo un legame lineare tra X e Y. Invece, *la indipendenza tra v.c. esclude ogni legame probabilistico* (cioè di qualsiasi tipo).

Allora, è bene sottolineare che:

- se non vi è alcun legame probabilistico (=*indipendenza*) tra X e Y, a maggior ragione, non vi potrà essere legame lineare;

- se non vi è legame lineare tra due v.c. (=*incorrelazione*) non è escluso che vi possano essere altri tipi di relazioni funzionali.

Esempio 9.13

Calcoliamo il coefficiente di correlazione lineare per la v.c. doppia (X, Y) introdotta nell'esempio 9.10. Per essa, risulta:

$$\mathbb{E}(X) = 6; \quad \mathbb{E}(Y) = \frac{4}{3}; \quad \mathbb{E}(XY) = 8;$$

per cui la covarianza è:

$$Cov(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y) = 8 - (6)\left(\frac{4}{3}\right) = 0;$$

Ne consegue che, essendo la covarianza nulla, sarà anche $Corr(X, Y) = 0$.

Pertanto, siamo in presenza di una v.c. doppia le cui componenti sono v.c. incorrelate (cioè la cui correlazione è 0) ma che *non sono indipendenti*. In sostanza, non vi è un legame di primo grado fra tali v.c. ma esiste invece un legame probabilistico tra X e Y perché, come abbiamo già discusso nell'esempio 9.10, la conoscenza di una delle due v.c. modifica la distribuzione di probabilità dell'altra.

□

Ambiente R

Variabili casuali

R9.1 Rappresentazione di una variabile casuale discreta

Le nozioni sviluppate nel Capitolo 9 non richiedono particolari elaborazioni numeriche per cui ci limitiamo ad utilizzare R per rappresentare graficamente le variabili casuali (v.c.) descritte nel testo.

La Figura 9.4 del testo mostra la distribuzione di probabilità e la funzione di ripartizione della v.c. generata dalla prova "lancio di un dado ben equilibrato" che assume probabilità costante e pari a 1/6; ad essa corrisponde una funzione di ripartizione a gradini di altezza costante e pari a tale probabilità.

Le istruzioni occorrenti per ottenere tale Figura sono le seguenti:

```
R==> m=6;  x=1:m; probx=rep(1/m,m);

R==> par(mfrow=c(1,2))  # suddividi l'area del grafico
                        # in 2 regioni

### Distribuzione di probabilità
R==> plot(x,probx, type="h",lwd=3, ylim=c(0,1.2*(1/m)),
       xlab="x",ylab="Pr(X=x)=1/6");
R==> points(x,probx,pch=19,cex=1.75);
abline(h=0);

### Funzione di ripartizione
R==> xx=0:(m+1); fripxx=c(0,cumsum(probx),1);
R==> plot(xx,fripxx, type="s",lwd=3, ylim=c(0,1),
       xlab="x", ylab="Pr(X<=x)");
R==> points(0:m,(0:m)/m,pch=19,cex=1.75);
R==> gradino=paste("1/",m,"",sep="");
R==> text((1:m)+0.4,(2*(1:m)-1)/(2*m),rep(gradino,m),cex=1);
R==> abline(h=0);

R==> par(mfrow=c(1,1))  # ripristina l'area del grafico
```

Si noti l'utilizzo della funzione grafica secondaria `points()` per inserire i punti all'apice delle barre che individuano le probabilità (le barre verticali sono ottenute con l'opzio-

ne `type="h"`) e all'inizio delle linee della funzione di ripartizione (il grafico a gradini è ottenuto con l'opzione `type="s"`).

Una caratteristica delle istruzioni sopra riportate è che esse sono completamente parametrizzate in funzione di m (che nel caso specifico è stato posto pari a $m = 6$). Quindi, come vedremo più avanti (paragrafo 10.5), esse possono essere utilizzate per rappresentare qualsiasi v.c. che abbia probabilità costante e pari a $1/m$, per un prefissato m.

R9.2 Rappresentazione di una variabile casuale continua

Se una v.c. continua ha funzione di densità $f(x)$ e funzione di ripartizione $F(x)$ definite, rispettivamente, da:

$$f(x) = \begin{cases} 2x, & \text{per ogni } 0 \le x \le 1; \\ 0, & \text{altrove.} \end{cases} \qquad F(x) = \begin{cases} 0, & -\infty < x \le 0; \\ x^2, & 0 < x \le 1; \\ 1, & 1 < x < +\infty; \end{cases}$$

la rappresentazione grafica di tali funzioni richiede il loro calcolo in un insieme di valori di x piuttosto ravvicinato per ottenere un grafico che generi un effetto "continuo". Nel seguito, abbiamo utilizzato $np = 1000$ punti, riproducendo in tal modo la Figura 9.5 del testo.

```
R==> np=1000;
R==> x=(0:np)/np;          fx=2*x;
R==> xx=(0:(1.5*np))/np;   fxx=x^2;
R==>                       fripx=c(fxx[1:(np+1)],rep(1,0.5*np));

R==> par(mfrow=c(1,2))   # suddividi l'area del grafico
                         # in 2 regioni

R==> ### Funzione di densità
R==> testofx=paste(expression(f(x)),expression(2*x),sep="=");
R==> plot(x,fx,type="l",lwd=2, xlim=c(0,1),
        ylim=c(0,2),xlab="x",ylab=testofx);
R==> segments(0,2,1,2,lty="dotted");
R==> segments(1,0,1,2,lty="dotted");
R==> segments(0,0,1,0,lty="solid");
R==> segments(0,0,0,2,lty="solid");

R==> ### Funzione di ripartizione
R==> testofripx=paste(expression(F(x)),expression(x^2),sep="=");
R==> plot(xx,fripx,type="l",lwd=2, xlim=c(0,1.5),ylim=c(0,1),
       xlab="x", ylab=testofripx);
R==> segments(0,1,1,1,lty="dotted");
R==> segments(1,0,1,1,lty="dotted");
R==> segments(0,0,1.5,0,lty="solid");
R==> segments(0,0,0,1,lty="solid");

R==> par(mfrow=c(1,1))   # ripristina l'area del grafico
```

Capitolo decimo

Principali modelli probabilistici

10.1 Introduzione

Anche se è praticamente impossibile classificare tutti i possibili esperimenti probabilistici, esistono numerosi elementi in comune tra quelli che usualmente si incontrano nelle applicazioni; il che consente di studiare alcuni schemi standard per le prove e, quindi, per le v.c. ad esse associate. Tali schemi si definiscono *modelli probabilistici* e saranno descritti mediante *famiglie parametriche di v.c.*

o Per esempio, la scelta di una pallina da un'urna, la scelta di un elettore di votare a favore o contro un candidato al ballottaggio, la scelta di un consumatore verso l'acquisto o meno di un prodotto, e così via, sono situazioni sperimentali molto differenti tra loro, ma che sono tutte schematizzabili mediante un modello probabilistico standard (come vedremo, del tipo Bernoulli) il cui studio e le cui proprietà possono essere adattate alle varie situazioni in funzione di una caratteristica numerica (detta *parametro*) che individua la prova specifica.

Formalmente, una ***famiglia parametrica di v.c.*** è un insieme di v.c. $X \sim f(x; \theta)$ descritte da un ***parametro*** θ (o da un vettore di parametri) che appartiene ad un insieme $\Omega(\theta)$, detto ***spazio parametrico***.

Con $f(x; \theta)$ intendiamo, per semplicità (qui e nel seguito, ove non diversamente definito) sia la distribuzione di probabilità – quando X è una v.c. discreta – che la funzione di densità – quando X è una v.c. continua. Per unificare i due casi, si potrebbe scrivere: $X \sim F(x; \theta)$ perché in entrambi i casi la funzione di ripartizione è l'elemento caratterizzante sia per v.c. discrete che continue.

In pratica, una famiglia parametrica di v.c. è una collezione di v.c. caratterizzate dal fatto che la loro *funzione di ripartizione* $X \sim F(x; \theta)$ ha la stessa forma funzionale e dipende da un parametro θ che specifica la singola v.c. appartenente alla famiglia predetta. Quindi, tali v.c. presentano la stessa forma della distribuzione di probabilità, se trattasi di *v.c. discrete*, e la stessa forma della funzione di densità, se trattasi di *v.c. continue*.

I membri della famiglia parametrica di v.c. (cioè, le singole v.c.) si distinguono esclusivamente per lo specifico valore numerico del parametro θ.

o Per comprendere meglio il concetto di «famiglia» parametrica di v.c. si pensi ad una famiglia umana i cui componenti, generalmente, assumono il cognome del capofamiglia (Rossi) ed i cui membri sono individuati dalla specificazione del nome proprio (Mario, Francesca, Giovanna, etc.). Ora, tutti i componenti della famiglia sono dei Rossi e la individuazione di uno di loro avviene grazie allo specifico «parametro» attribuito a ciascuno, cioè il nome.

o I principali modelli probabilistici che studieremo nel seguito sono quelli più semplici ed anche più diffusi nelle applicazioni, e saranno presentati prima quelli che riguardano v.c. continue e poi quelli riferiti a v.c. discrete (sul piano espositivo, ciò facilita la discussione che faremo delle approssimazioni di alcune v.c. discrete tramite una particolare v.c. continua).

o Sul piano della notazione, indicheremo con θ il parametro (o i parametri) che caratterizzano le v.c. che studieremo, ma preferiremo usare altri simboli quando, nella letteratura statistica, vi è uniformità di simbologia per tali famiglie parametriche.

10.2 La variabile casuale Uniforme continua

Il modello probabilistico che genera la v.c. Uniforme continua è utilizzato per la semplicità della sua formulazione analitica ma, in effetti, esso si incontra quasi sempre, in situazioni artificiali, come giochi, esperimenti geometrici, e così via.

o Tale v.c. si può far derivare da una prova il cui meccanismo probabilistico (ago ruotante, rotolamento di una pallina perfettamente sferica e senza attrito in un astuccio sino a fermarsi, etc.) conduce alla determinazione di un numero reale che, per motivi fisici, è compreso in un intervallo delimitato. Ragioni di simmetria e indifferenza implicano che sotto-intervalli di pari ampiezza abbiano la stessa probabilità e, di conseguenza, la probabilità che la prova generi un numero reale compreso in un certo intervallo deve essere proporzionale alla lunghezza dell'intervallo medesimo.

Definiamo X una ***v.c. Uniforme continua*** definita sul supporto (a, b), e scriveremo $X \sim U(a, b)$, se la funzione di densità di tale v.c. è:

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \le x \le b; \\ 0, & \text{altrove.} \end{cases}$$

Tale modello probabilistico implica che ogni intervallo reale deve avere una probabilità di verificarsi proporzionale alla sua lunghezza.

Mediante semplici passaggi, si derivano i *momenti caratteristici* della v.c. Uniforme continua che risultano essere:

$$\mathbb{E}(X) = \frac{a+b}{2};\ Var(X) = \frac{(b-a)^2}{12};\ Asym(X) = 0;\ Kurt(X) = 1.8\,.$$

Il valore medio e la mediana di tale v.c. coincidono con il valore collocato al centro dell'intervallo di definizione; essa è simmetrica e platicurtica (cioè con distribuzione molto piatta), senza alcuna moda.

Ricordando che la funzione di ripartizione è definita da: $F(x) = \int_{-\infty}^{x} f(w)dw$ si ottiene:

$$F(x) = \begin{cases} 0, & \text{per } x \le a; \\ \dfrac{x-a}{b-a}, & a \le x \le b; \\ 1, & \text{per } x \ge b. \end{cases}$$

Come si vede, essa è nulla per valori dell'argomento inferiori al minimo, vale 1 per valori dell'argomento superiori al massimo e cresce linearmente sul supporto di definizione della v.c. Uniforme continua.

o La trasformazione lineare di una v.c. Uniforme continua genera un'altra v.c. Uniforme, ma la combinazione lineare di v.c. Uniformi indipendenti genera v.c. le cui funzioni di densità sono molto diverse.

Così, la somma di due v.c. Uniformi continue genera una v.c. la cui funzione di densità consiste di due rette di pendenza opposta che si incontrano nel punto di ascissa pari al valore medio della somma (in tale caso, la v.c. si chiama *triangolare*); la somma di tre v.c. Uniformi continue genera una v.c. la cui funzione di densità è la giustapposizione di tre archi di parabole; in generale, la somma S_n di n v.c. Uniformi continue è una v.c. la cui funzione di densità è la giustapposizione di n polinomi di grado $(n-1)$, ciascuno definito su sotto-intervalli di uguale ampiezza.

Al crescere del numero n di addendi, la funzione di densità della v.c. S_n standardizzata, tende velocemente ad assumere una forma standard campanulare e simmetrica; per esempio, se il supporto delle v.c. addendi è $(0, 1)$ la funzione di densità della v.c. somma standardizzata, già quando $n = 12$, è molto ben approssimata dalla funzione di densità: $f(z) = \left(1/\sqrt{\pi n/6}\right) \exp\{-6(z - n/2)^2\}$. Questa situazione sarà ampiamente discussa e commentata nell'esempio 11.1.

Esempio 10.1

Un esempio di prova che genera una v.c. Uniforme continua deriva dal cosiddetto «ago ruotante». La prova consiste nel far girare, su un piano orizzontale, un ago che ruota attorno ad un centro descrivendo una circonferenza, dopo aver impresso ad esso una sufficiente spinta in modo che non sia prevedibile il suo punto di arresto.

Se si indica con X l'angolo che la freccia, fermandosi, forma con una origine arbitraria (come indicato in Figura 10.1), allora la prova genera un numero reale che è l'angolo compreso tra 0° e 360°, se misurato in gradi, oppure tra 0 e 2π, se misurato in radianti.

In tale prova, se il piano su cui è posizionato il perno non è inclinato, se l'attrito è uniforme e se la spinta che viene data è tale da non privilegiare alcun risultato, è ragionevole ammettere che ogni arco di pari ampiezza possieda pari probabilità di verificarsi.

Per esempio, la probabilità che la freccia si fermi in uno dei 4 quadranti deve essere 1/4; che si fermi in uno degli 8 ottanti deve essere 1/8; ... Sono quindi rispettate le ipotesi che sono alla base della v.c. Uniforme continua per cui tale prova genera (se gli angoli sono misurati in radianti) la v.c. $U \sim U(0, 2\pi)$ la cui funzione di densità è:

$$f(x) = \begin{cases} \dfrac{1}{2\pi}, & 0 \le x \le 2\pi; \\ 0, & \text{altrove.} \end{cases}$$

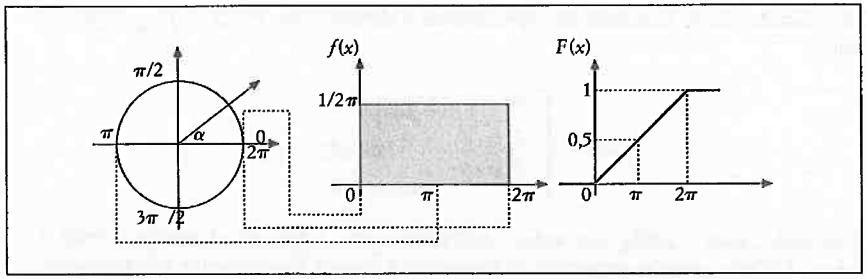

FIGURA 10.1. Variabile casuale Uniforme continua associata alla prova «Ago ruotante».

Da essa, si derivano sia la funzione di ripartizione (di cui alla figura precedente) che i momenti caratteristici (basta porre: $a = 0$; $b = 2\pi$, nelle formule generali prima esposte). Sono altresì ovvie le variazioni da apportare quando gli angoli sono misurati in gradi ($a = 0$; $b = 360$).

□

10.3 La variabile casuale Normale

La v.c. Normale è di gran lunga la distribuzione più importante del Calcolo delle probabilità e della Statistica, al punto che appare difficile immaginare lo sviluppo di tali discipline senza riconoscere ad essa un ruolo centrale. La sua stessa denominazione (*normale* suggerisce che essa è la «norma»!) esplicita che la sua presenza nelle applicazioni e negli sviluppi teorici è quasi una regola. In effetti, la v.c. Normale approssima la distribuzione empirica di moltissimi fenomeni reali, ma è anche punto di riferimento per stabilire confronti, dedurre risultati asintotici e controllare allontanamenti dalla distribuzione di tale v.c.

o La sua introduzione nella storia della probabilità è legata alla ricerca della distribuzione degli errori accidentali. Galileo descrisse con precisione, sin dal 1632, le caratteristiche essenziali che una distribuzione degli errori doveva possedere nel «*Dialogo dei massimi sistemi*» ma non si preoccupò di derivarne la forma analitica.

o La prima apparizione della funzione di densità della v.c. Normale avviene nel 1733 ad opera di De Moivre che la utilizza come approssimazione alla somma di v.c. Binomiali (che esamineremo nel paragrafo 10.7). Negli anni 1770-71 Daniel Bernoulli fornisce la prima tavola della funzione di densità mentre Laplace a partire dal 1810 – in una serie di lavori che saranno fondamentali per lo sviluppo del Calcolo delle probabilità e della Statistica – la ritrova nei Teoremi Limite Centrale (paragrafo 11.3).

o Nel 1809 Gauss pubblica la «*Theoria Motus Corporum Coelestium*» dove utilizza la v.c. Normale nell'ambito del principio dei minimi quadrati, introdotto da Legendre nel 1805, e del metodo della massima verosimiglianza. Gauss in tale opera dichiara di aver utilizzato questi risultati sin dal 1795. Questa affermazione, e l'autorità indiscussa di Gauss nell'ambito delle scienze matematiche, hanno indotto a definire la *v.c. Normale* come *v.c. di Gauss* o *v.c. Gaussiana.*

o Grazie ai Teoremi Limite Centrale, e alle interpretazioni estensive che ne furono date, la v.c. Normale non fu applicata alle sole misurazioni fisiche. Quételet nel 1846 la utilizzò per studi antropometrici, Maxwell nel 1860 per la teoria cinetica dei gas, Galton nel 1889 per lo studio dei fenomeni naturali e biologici. Quindi, a partire dagli inizi del 1900, Student, Pearson e Fisher posero le basi inferenziali della Statistica mediante il campionamento, la stima e il test soprattutto per popolazioni estratte da v.c. Normali e derivarono quelle distribuzioni notevoli, esatte o asintotiche, che oggi utilizziamo.

o A titolo di curiosità, si può ricordare che la funzione di densità della v.c. Normale è l'unica formula matematica che, nella storia, sia stata impressa su una banconota con ampia circolazione (i 10 marchi tedeschi emessi nel 1994 e circolati sino all'introduzione dell'Euro nel 2002).

Una v.c. X continua si dice *v.c. Normale* (oppure *v.c. Gaussiana*) con parametri μ e σ^2, e la si indica con $X \sim N(\mu, \sigma^2)$, se è definita su tutto l'asse reale con funzione di densità:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}, \quad -\infty < x < \infty.$$

Poiché il parametro μ può assumere qualsiasi valore reale mentre il parametro σ^2 può solo essere non-negativo, lo spazio parametrico di questa v.c. è il semipiano positivo, cioè:

$$\Omega(\mu, \sigma^2) = \{(\mu, \sigma^2) : -\infty < \mu < +\infty;\ 0 \leq \sigma^2 < +\infty\}.$$

Lo studio analitico della funzione di densità di una v.c. Normale (Figura 10.2) mostra che essa ha una forma campanulare simmetrica rispetto al suo valore medio (punto di ascissa $x = \mu$), in corrispondenza del quale si presenta il massimo di $f(x)$. Quindi, il parametro μ è la moda, la mediana e il valore medio della v.c. X.

Utilizzando le derivate seconde, si dimostra che $f(x)$ presenta due flessi equidistanti dal punto $x = \mu$ in corrispondenza delle ascisse $x = \mu \pm \sigma$. Inoltre, $f(x) \to 0$ per $x \to \pm\infty$, cioè l'asse x è un asintoto orizzontale per tale funzione.

Se – a parità di σ^2 – si modifica il valore medio μ, allora la funzione di densità della v.c. Normale subisce una traslazione lungo l'asse x. Invece, al crescere di σ^2, – a parità di μ – i flessi si allontanano da μ e la funzione di densità attribuisce maggiore probabilità a valori più lontani dal valore medio μ; per contro, al diminuire di σ^2 – a parità di μ – i flessi si avvicinano al valore medio μ ed aumenta la probabilità che la v.c. assuma valori attorno al valore medio μ. Infine, quando $\sigma^2 \to 0$, la distribuzione della v.c. Normale tende a divenire degenere perché assume, con probabilità 1, valori infinitamente vicini ad $x = \mu$. Tali situazioni sono schematizzate nella Figura 10.3.

I momenti caratteristici della v.c. Normale (di difficile derivazione analitica) risultano uguali a:

$$\boxed{\mathbb{E}(X) = \mu;\ Var(X) = \sigma^2;\ Asym(X) = 0;\ Kurt(X) = 3.}$$

o Grazie a tali caratteristiche, spesso si afferma che se una v.c. presenta una forma simmetrica e una curtosi vicina a 3 essa può essere approssimata da una v.c. Normale di uguali valore medio e varianza. In effetti, tale approssimazione non è rigorosa ma, *se la v.c. presenta una sola moda*, essa può essere applicata nella maggior parte dei casi.

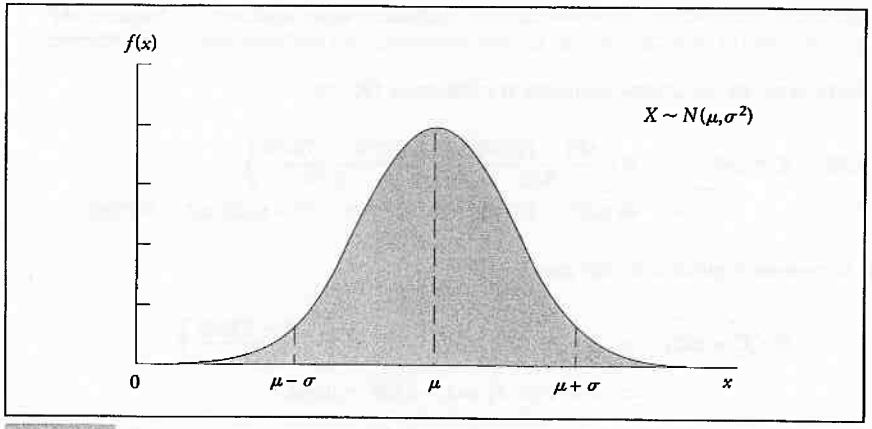

FIGURA 10.2. Funzione di densit della variabile casuale Normale.

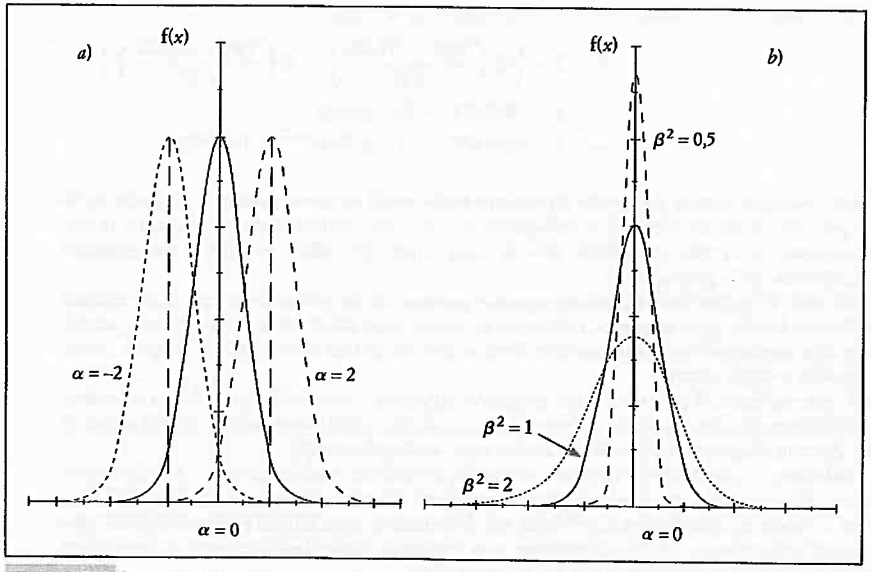

FIGURA 10.3. Funzioni di densità della variabile casuale Normale, al variare: *a)* del valore medio; *b)* della varianza.

Se $X \sim N(\mu, \sigma^2)$, la trasformazione lineare $Z = \dfrac{X-\mu}{\sigma}$ definisce la ***v.c. Normale standardizzata*** $Z \sim N(0,1)$, le cui funzioni di densità e di ripartizione sono universalmente indicate con i simboli $\phi(z)$ e $\Phi(z)$, rispettivamente. Tali funzioni sono espresse, quindi, da:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}; \quad \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}w^2} dw; \quad -\infty < z < \infty.$$

Ovviamente, come *tutte* le v.c. standardizzate (paragrafo 9.6), anche la v.c. Normale standardizzata possiede valore medio uguale a 0 e varianza uguale ad 1.

La funzione di ripartizione della v.c. Z non ammette una formulazione esplicita per cui è stato necessario predisporre delle tavole che, per opportuni valori di z, forniscano l'integrale $\Phi(z)$ con sufficiente accuratezza: alla fine del volume, è riportata nella Tavola 1 la funzione $\Phi(z)$ per i più rilevanti valori di $z \geq 0$. Peraltro, la funzione di ripartizione $\Phi(z)$ è presente nella gran parte dei software statistici e dei linguaggi di programmazione (matematici e statistici).

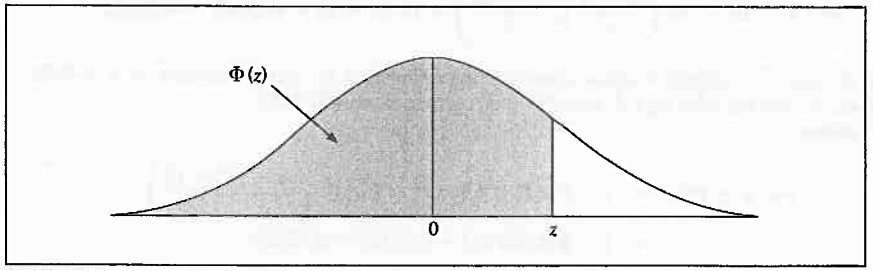

FIGURA 10.4. Funzione di ripartizione della variabile casuale Normale standardizzata.

Sfruttando la simmetria della funzione di densità, è necessario conoscere $\Phi(z)$ per i soli valori di $z \geq 0$. Infatti, $\Phi(0) = Pr(X \leq 0) = 0.5$ ed inoltre (come è agevole dedurre dalla Figura 10.4):

$$\Phi(-z) = 1 - \Phi(z), \ \forall z \geq 0.$$

La conoscenza della funzione di ripartizione della v.c. $Z \sim N(0,1)$ è sufficiente per calcolare la probabilità di eventi che riguardano una *qualsiasi* v.c. $X \sim N(\mu, \sigma^2)$, mediante una semplice trasformazione.

Infatti, sono immediati i seguenti passaggi:

$$\begin{aligned} Pr(x_0 < X \leq x_1) &= Pr\left(\frac{x_0-\mu}{\sigma} < \frac{X-\mu}{\sigma} \leq \frac{x_1-\mu}{\sigma}\right) \\ &= Pr\left(\frac{x_0-\mu}{\sigma} < Z \leq \frac{x_1-\mu}{\sigma}\right), \end{aligned}$$

per cui, ricordando la definizione della funzione di ripartizione $\Phi(z)$ della v.c. $Z \sim N(0,1)$, si ottiene la seguente formula valida per ogni v.c. $X \sim N(\mu, \sigma^2)$:

$$P_r(x_0 < X \leq x_1) = \Phi\left(\frac{x_1 - \mu}{\sigma}\right) - \Phi\left(\frac{x_0 - \mu}{\sigma}\right).$$

Tale espressione mostra che *per calcolare la probabilità che una v.c. Normale assuma valori in un qualsiasi intervallo basta standardizzare gli estremi dell'intervallo* ed utilizzare, poi, le tavole della funzione di ripartizione $\Phi(z)$.

Esempio 10.2

Calcoliamo la probabilità di alcuni eventi relativi ad una prefissata v.c. $X \sim N(10, 4)$, per la quale quindi i parametri sono: $\mu = 10; \sigma^2 = 4$ (quindi, $\sigma = 2$).

Per esempio:

$$P_r(X \leq 16) = P_r\left(\frac{X - \mu}{\sigma} \leq \frac{16 - 10}{2}\right) = P_r(Z \leq 3) = \Phi(3.00) = 0.99865;$$

dove il valore di 0.99865 è stato ottenuto dalla Tavola 1 in corrispondenza di $z = 3.00$ (quindi, in pratica, alla riga di $z = 3.00$ e sotto la colonna di 0.00).

Inoltre,

$$\begin{aligned} P_r(X > 12) &= 1 - P_r(X \leq 12) = 1 - P_r\left(\frac{X - \mu}{\sigma} \leq \frac{12 - 10}{2}\right) \\ &= 1 - \Phi(1.00) = 1 - 0.84134 = 0.15866; \end{aligned}$$

$$\begin{aligned} P_r(12 < X \leq 16) &= \Phi\left(\frac{16 - 10}{2}\right) - \Phi\left(\frac{12 - 10}{2}\right) \\ &= \Phi(3) - \Phi(1) = 0.99865 - 0.84134 = 0.15731; \end{aligned}$$

$$P_r(X \leq 8) = P_r\left(\frac{X - \mu}{\sigma} \leq \frac{8 - 10}{2}\right) = \Phi(-1) = 1 - \Phi(1) = 0.15866;$$

$$\begin{aligned} P_r(8 < X \leq 16) &= \Phi\left(\frac{16 - 10}{2}\right) - \Phi\left(\frac{8 - 10}{2}\right) \\ &= \Phi(3) - \Phi(-1) = 0.99865 - (1 - 0.84134) = 0.83999. \end{aligned}$$

□

Esempio 10.3

Supponiamo che la v.c. $X =$ «Altezza» (in cm) di una persona adulta di sesso maschile, estratta a caso da una popolazione molto ampia, sia ben approssimata da una v.c. Normale con parametri $\mu = 176.25; \sigma^2 = 40.34752$ (e, quindi, $\sigma \simeq 6.35$).

Calcoliamo, nell'ordine, la probabilità dei seguenti eventi mediante la formula che utilizza la funzione di ripartizione degli estremi standardizzati dell'intervallo di riferimento:

i) la persona ha un'altezza compresa tra 170 cm e 185 cm;

$$\begin{aligned} P_r(170 \leq X < 185) &= \Phi\left(\frac{185 - 176.25}{6.35}\right) - \Phi\left(\frac{170 - 176.25}{6.35}\right) \\ &= \Phi(1.38) - \Phi(-0.98) = 0.91621 - (1 - 0.83646) = 0.75267. \end{aligned}$$

ii) la persona è più alta di 190 cm;

$$\begin{aligned} P_r(X > 190) &= 1 - P_r(X \leq 190) = 1 - \Phi\left(\frac{190 - 176.25}{6.35}\right) \\ &= 1 - \Phi(2.17) = 1 - 0.985 = 0.015. \end{aligned}$$

iii) la persona è più bassa di 160 cm oppure più alta di 200 cm;

$$\begin{aligned} P_r((X < 160) \cup (X > 200)) &= 1 - P_r(160 < X < 200) \\ &= 1 - \left(\Phi\left(\frac{200 - 176.25}{6.35}\right) - \Phi\left(\frac{160 - 176.25}{6.35}\right)\right) \\ &= 1 - (\Phi(3.74) - \Phi(-2.56)) \\ &= 1 - (0.99991 - (1 - 0.99477)) = 0.00532. \end{aligned}$$

Questo esempio deriva da quelle situazioni nelle quali occorre valutare la taglia delle persone per offrire un prodotto o predisporre un servizio, individuale o collettivo (tute, divise, automobili, arredo e mobilio per la casa e per gli uffici, panchine nei giardini pubblici, altezza delle porte, etc.).

In tali casi, è importante conoscere quante persone si presenteranno con determinate caratteristiche fisiche (per esempio, l'altezza) in modo da soddisfare la richiesta con scorte adeguate alla domanda o comunque con beni o servizi idonei a soddisfare la gran parte della clientela o degli utenti.

Così, per esempio, il primo quesito proposto riguarda, verosimilmente, le caratteristiche più diffuse nella popolazione; invece, il terzo quesito può interessare i distributori di area che devono disporre del prodotto anche per i casi più estremi.

Per valutare la qualità della risposta numerica fornita da questo esempio, abbiamo utilizzato nei calcoli i corrispondenti valori osservati per i Maschi del data set STUDENTI sostituendo ai valori dei parametri μ, σ^2 della v.c. Normale le quantità 176.25 e 40.34752 (che sono, rispettivamente, la media aritmetica e la varianza della distribuzione di frequenza delle altezze rilevate sugli studenti di sesso maschile).

La correttezza e la bontà di tale procedura sarà apprezzata solo dopo lo studio dell'Inferenza; tuttavia, lo scopo e la discussione di questo esempio sono limitati all'uso della v.c. Normale. La tabella seguente confronta le frequenze relative degli studenti che hanno la caratteristica individuata con la probabilità dell'evento corrispondente:

Modalità // Eventi	*Frequenza relativa*	*Probabilità*
$(170 \leq X < 185)$	0.773	0.753
$X > 190$	0.014	0.015
$(X < 160) \cup (X > 200)$	0.003	0.005

Tale confronto mostra la utilità del modello probabilistico Normale per determinare l'ordine di grandezza dell'occorrenza degli eventi indicati, anche in una situazione in cui le misurazioni non sono reali ma *dichiarate dagli studenti*, quindi con inevitabili arrotondamenti ed imprecisioni.

□

Poiché non è esplicitabile in forma chiusa la funzione di ripartizione della v.c. Z, non lo saranno nemmeno i quantili z_α definiti da: $Pr(Z \leq z_\alpha) = 1 - \alpha$, ovvero le quantità z_α tali che vi sia una probabilità pari ad $1 - \alpha$ che esse non siano superate.

Tali quantità saranno frequentemente utilizzate nell'Inferenza statistica per cui sono riportate in fondo alla Tavola 1. Per esempio, risulta che:

$$z_{0.05} = 1.64485; \quad z_{0.01} = 2.32635; \quad z_{0.001} = 3.09023;$$

perché per la v.c. Normale standardizzata $Z \sim N(0,1)$, si ha rispettivamente:

$$Pr(Z \leq 1.64485) = 0.95; \; Pr(Z \leq 2.32635) = 0.99; \; Pr(Z \leq 3.09023) = 0.999\,.$$

Per la v.c. Normale, le probabilità che si verifichino valori compresi in intervalli centrati sul valore medio μ e da esso distanti per multipli di σ sono delle costanti che *non* dipendono dal valore medio e dalla varianza. Poiché il valore medio è anche la moda della v.c., tali intervalli sono quelli che a parità di ampiezza racchiudono la maggior parte della probabilità di tale distribuzione.

A tal fine, per qualsiasi v.c. $X \sim N(\mu, \sigma^2)$, è utile conoscere i seguenti valori:

$$\begin{aligned} Pr(\mu - \sigma \leq X \leq \mu + \sigma) &= 0.68269 \\ Pr(\mu - 2\sigma \leq X \leq \mu + 2\sigma) &= 0.95450 \\ Pr(\mu - 3\sigma \leq X \leq \mu + 3\sigma) &= 0.99730 \\ Pr(\mu - 4\sigma \leq X \leq \mu + 4\sigma) &= 0.99994 \end{aligned}$$

Come si vede, la gran parte (il 95%) e la quasi totalità (il 99.7%) dei valori reali che una v.c. Normale può assumere sono compresi tra $\mu \pm 2\sigma$ e tra $\mu \pm 3\sigma$, rispettivamente.

Ciò ha suggerito, per fenomeni che si possono concepire come realizzazioni di v.c. Normali, di considerare *atipici* o *anomali* quei valori che sono più lontani dal valore medio del doppio o del triplo dello scarto quadratico medio. Inoltre, le costanti fisiche vengono spesso riportate con una indicazione dello scarto quadratico medio σ, in modo da fornire una idea circa la collocazione entro cui può oscillare il valore della grandezza da misurare.

Un'ulteriore importante proprietà della famiglia delle v.c. Normali (che ci limitiamo ad enunciare e di cui vedremo qualche applicazione in ambito inferenziale) è un teorema chiamato ***proprietà riproduttiva della v.c. Normale***:

La combinazione lineare di v.c. Normali e indipendenti è ancora una v.c. Normale, che ha per valore medio la combinazione lineare dei valori medi e per varianza la combinazione lineare delle varianze con i quadrati dei coefficienti.

Sul piano formale, questo teorema afferma che se $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2, \ldots, n$ è una sequenza di v.c. indipendenti, allora per qualsiasi sequenza di costanti reali $a_1, a_2, \ldots, a_n$, la v.c. combinazione lineare sarà:

$$\sum_{i=1}^{n} a_i X_i \sim N(\mu, \sigma^2), \quad \text{dove:} \quad \mu = \sum_{i=1}^{n} a_i \mu_i; \; \sigma^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2.$$

10.4 Variabili casuali connesse alla variabile casuale Normale

In questo paragrafo, definiamo alcune v.c. continue connesse e/o derivabili dalla v.c. Normale standardizzata. Esse sono di fondamentale importanza per la teoria dell'Inferenza statistica, soprattutto con riferimento ai quantili necessari nell'ambito dei test delle ipotesi.

∘ Date le finalità ed il livello di questo volume, ci limitiamo ad una schematica definizione e ad una illustrazione essenziale delle loro principali proprietà. Come si vedrà, esse sono tutte descritte da uno o due parametri che – per le connessioni con i successivi problemi inferenziali – vengono chiamati *gradi di libertà.*

• La somma di g v.c. Normali standardizzate indipendenti e al quadrato è una v.c. continua sul supporto $(0, +\infty)$ definita ***v.c. Chi-quadrato*** e indicata con $X \sim \chi^2_{(g)}$. Essa è una famiglia parametrica di v.c. i cui membri sono caratterizzati dal parametro g, detto *gradi di libertà della v.c. Chi-quadrato.*

La funzione di densità di tale v.c. è asimmetrica positiva e la funzione di ripartizione è derivabile solo mediante approssimazioni numeriche. La Tavola 2 in Appendice riporta i quantili $\chi^2_{(g;\,\alpha)}$ della v.c. $X \sim \chi^2_{(g)}$ al variare dei gradi di libertà g e per prefissati valori di $1 - \alpha$, definiti dalla relazione:

$$Pr\left(X \leq \chi^2_{(g;\,\alpha)}\right) = 1 - \alpha.$$

∘ Per esempio, se $g = 10$, dalla Tavola 2 si deduce:

$$\chi^2_{(g;\,0.05)} = 18.307; \; \chi^2_{(g;\,0.01)} = 23.209;$$

perché risulta, rispettivamente:

$$Pr(X \leq 18.307) = 1 - 0.05 = 0.95; \; Pr(X \leq 23.209) = 1 - 0.01 = 0.99\,.$$

I momenti caratteristici della v.c. Chi-quadrato con g gradi di libertà sono:

$$\boxed{\mathbb{E}(X) = g; \; Var(X) = 2g; \; Asym(X) = \sqrt{\tfrac{8}{g}}; \; Kurt(X) = 3 + \tfrac{12}{g}.}$$

Da essi si deduce che la funzione di densità è asimmetrica positiva ma, al crescere di g, tende ad assumere una forma sempre più vicina alla simmetria e alla distribuzione Normale.

• Una v.c. immediatamente connessa alla precedente è la v.c. di Fisher (detta anche di Fisher-Snedecor) definita come il rapporto tra due v.c. Chi-quadrato indipendenti tra loro e divise per i rispettivi gradi di libertà.

Formalmente, se $X_1 \sim \chi^2_{(g_1)}$ e $X_2 \sim \chi^2_{(g_2)}$ sono due v.c. Chi-quadrato indipendenti si definisce ***v.c. F di Fisher***, e la indicheremo con $X \sim F_{(g_1,g_2)}$, la seguente v.c.:

$$X = \frac{X_1/g_1}{X_2/g_2}.$$

I due parametri g_1, g_2 che caratterizzano tale v.c. sono definiti, rispettivamente, *gradi di libertà del numeratore* e *gradi di libertà del denominatore*.

Dalla definizione, è agevole dedurre che se X è una v.c. di Fisher lo è anche la v.c. $1/X$ con i gradi di libertà scambiati tra loro.

Di tale v.c. sono necessari i quantili $F_{(g_1,g_2;\alpha)}$ che, con ragionamento simile al precedente, sono disponibili sulla Tavola 4 in Appendice per i soli valori di $\alpha = 0.05$ e $\alpha = 0.01$, rispettivamente, essendo questi quelli più utilizzati nella teoria del test delle ipotesi.

◦ Per esempio, se $g_1 = 8; g_2 = 5$, dalla Tavola 4 si deduce:

$$F_{(8,5;\,0.05)} = 4.82;\ F_{(8,5;\,0.01)} = 10.27;$$

perché per la v.c. $X \sim F_{(8,5)}$ risulta, rispettivamente:

$$P_r(X \leq 4.82) = 1 - 0.05 = 0.95;\ P_r(X \leq 10.27) = 1 - 0.01 = 0.99.$$

I momenti caratteristici di tale v.c. sono espressioni complesse dei parametri e, in ogni caso, derivano da una funzione di densità asimmetrica positiva con una asimmetria che si riduce al crescere dei gradi di libertà; tuttavia, la funzione di densità di tale v.c. tende a quella di una v.c. Normale solo se entrambi i gradi di libertà tendono all'infinito.

• Un'altra v.c. di fondamentale importanza per l'Inferenza statistica fu derivata dal chimico W.S. Gosset, che firmava i suoi lavori con lo pseudonimo di «Student», per cui oggi essa è universalmente chiamata v.c. di Student.

Formalmente, se $Z \sim N(0,1)$ è una v.c. Normale standardizzata e indipendente da una v.c. $Y \sim \chi^2_{(g)}$, allora si definisce ***v.c. di Student*** e la si indica con $X \sim \mathcal{T}_{(g)}$ la seguente v.c.:

$$X = \frac{Z}{\sqrt{\frac{Y}{g}}},$$

ottenuta facendo il rapporto tra una v.c. Normale standardizzata Z e la radice quadrata di una v.c. Chi-quadrato Y, indipendente dal numeratore, divisa per i suoi gradi di libertà. È agevole convincersi che il quadrato di una v.c. di Student con g gradi di libertà è una v.c. di Fisher nella quale $g_1 = 1;\ g_2 = g$.

Tale famiglia parametrica di v.c. continue è definita sull'intero asse reale (cioè assume tutti i valori reali) ed è specificata dal parametro g, detto *gradi di libertà* della v.c. di Student.

La funzione di densità della v.c. di Student è sempre simmetrica e di valore medio 0 ed assume una forma molto simile a quella della v.c. Normale standardizzata alla quale tende assai velocemente al crescere dei gradi di libertà.

Tuttavia, soprattutto per valori di g piccoli o moderati, si caratterizza per una curtosi leggermente più elevata e per code più pesanti della v.c. Normale. Grazie a tale comportamento, la v.c. di Student viene utilizzata nelle situazioni nelle quali i fenomeni reali assumono valori estremi con frequenza maggiore di quanto sia prevedibile mediante la v.c. Normale (ciò accade per alcune variabili di interesse finanziario, per esempio).

Al crescere del parametro g, la v.c. di Student è molto ben approssimata dalla v.c. Normale standardizzata, per cui l'uso della Tavola 3 in Appendice per il calcolo dei quantili $\mathcal{T}_{(g;\alpha)}$ è utile solo quando $g < 30$.

◦ Per esempio, se $g = 12$, dalla Tavola 3 si deduce che se $X \sim \mathcal{T}_{(12)}$, allora:

$$\mathcal{T}_{(12;\,0.05)} = 1.782;\ \mathcal{T}_{(12;\,0.001)} = 3.930;$$

perché risulta, rispettivamente:

$$P_r(X \leq 1.782) = 1 - 0.05 = 0.95;\quad P_r(X \leq 3.930) = 1 - 0.001 = 0.999.$$

10.5 La variabile casuale Uniforme discreta

La prova che genera una v.c. Uniforme discreta si può assimilare all'estrazione di una pallina da un'urna che contiene m palline identiche ma numerate da 1 a m. Il numero intero $m \geq 1$ è il *parametro* che caratterizza tale v.c. discreta.

Formalmente, nell'estrazione casuale di una pallina da un'urna che contiene palline numerate da 1 a m, se all'evento «si è verificata la pallina che reca il numero x» si associa il numero reale x, per $x = 1, 2, \ldots, m$, allora si definisce la v.c. X ***Uniforme discreta*** (detta, a volte, *Rettangolare*) che assume i valori $x = 1, 2, \ldots, m$ con probabilità costante, cioè:

$$P_r(X = x) = \frac{1}{m},\quad \forall x = 1, 2, \ldots, m.$$

Tale v.c. sarà indicata con $X \sim Ud(m)$.

Essa è ben definita perché sono rispettate le due condizioni:

$$P_r(X = x) = \frac{1}{m} > 0;\quad \sum_{x=1}^{m} P_r(X = x) = \sum_{x=1}^{m} \frac{1}{m} = 1.$$

o Particolarizzando il valore del parametro m, si riconoscono come appartenenti a tale famiglia numerose v.c. discrete connesse a giochi molto comuni: se $m = 2$ siamo di fronte al «Lancio di una moneta corretta»; se $m = 6$ trattasi del «Lancio di un dado simmetrico e ben equilibrato»; se $m = 90$ è «L'estrazione del primo estratto nel gioco del Lotto», e così via.

o Tale prova è inoltre diffusa nei sondaggi di opinione, negli esperimenti di laboratorio, e così via.

La rappresentazione grafica di tale distribuzione è ovviamente costante (la Figura 10.5 illustra il caso di $m = 10$). La funzione di ripartizione è nulla in $(-\infty, 1)$; è una funzione a gradini di altezza costante e pari a $1/m$ in corrispondenza di ogni valore intero $1 \leq x \leq m$; vale 1 nell'intervallo $[m, +\infty)$.

A tale riguardo, si osservi che nell'esempio 9.5 (Figura 9.4) si è già studiata la distribuzione di probabilità e la funzione di ripartizione della v.c. $X \sim Ud(6)$.

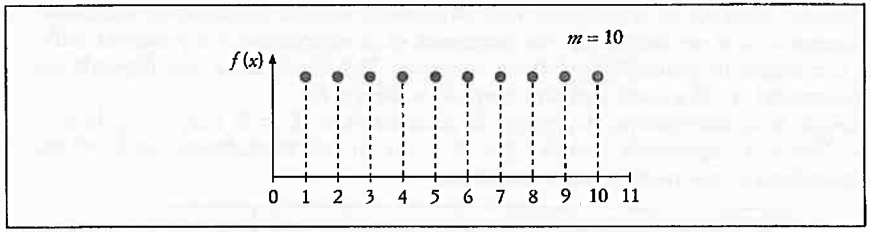

FIGURA 10.5. Distribuzione di probabilit della variabile casuale Uniforme discreta ($m = 10$).

In generale, quindi, la funzione di ripartizione di tale v.c. è così definita:

$$F(x) = \begin{cases} 0, & \text{per } x < 1; \\ \dfrac{k}{m}, & \text{per } k \leq x \leq k+1; \\ 1, & \text{per } x \geq m. \end{cases}$$

La v.c. $X \sim Ud(m)$ non presenta moda. Inoltre, i suoi momenti caratteristici risultano essere:

$$\mathbb{E}(X) = \frac{m+1}{2};\ Var(X) = \frac{m^2-1}{12};\ Asym(X) = 0;\ Kurt(X) = \frac{3\,(3m^2-7)}{5\,(m^2-1)}.$$

Da essi si deduce che il valore medio della v.c. Uniforme discreta coincide con il valore al centro del suo supporto e con la mediana; la distribuzione è simmetrica e la curtosi (definita per $m > 2$) tende velocemente al valore 1.8, confermando così di essere in presenza di una distribuzione completamente piatta.

10.6 La variabile casuale di Bernoulli

La v.c. di Bernoulli trae origine da una prova nella quale interessa esclusivamente stabilire se un evento E si è verificato oppure no (in tal caso, si verifica la sua negazione $\overline{E}$).

o Sono molteplici le situazioni reali in cui tale modello si incontra, perché tutte le prove il cui risultato è dicotomico possono ricondursi ad una v.c. di Bernoulli: il verificarsi di «Testa» nel lancio di una moneta; la scelta di una pallina bianca da un'urna che contiene palline bianche e nere; il verificarsi di un numero di punti superiore a 5 nel lancio di un dado oppure il verificarsi di un «doppio 6» nel lancio di due dadi; un tasso di inflazione nel 2005 inferiore al 2%; una precipitazione media settimanale in un certo bacino idrografico superiore ai $200\,mm/cm^2$; un esito favorevole ad un referendum abrogativo; un test negativo nell'esame di un individuo scelto a caso per controllare una malattia infettiva; il superamento dell'esame di Statistica; se il numero di auto in attesa al casello supera le 30 unità per minuto; la vincita di un concorso pubblico che prevede una molteplicità di prove; un esame clinico per il quale si controlla solo se la persona ha un livello di colesterolo giudicato normale; il sopravvivere di un assicurato sino ad una età prefissata; e così via.

o Tra le situazioni esemplificate, alcune sono strutturalmente dicotomiche (Testa o Croce, vincita o perdita, promosso o respinto, etc.); altre, lo diventano grazie alla dicotomizzazione di una prova il cui spazio di probabilità può essere concettualmente discreto o continuo, ma di cui interessa solo il superamento o meno di una soglia fissata per la prova (se il livello di colesterolo supera il livello normale; se la durata di vita supera una certa età; se il numero di auto in attesa supera 30 unità per minuto; se le precipitazioni medie settimanali superano i 200 mm/cm^2, etc.).

o In tali contesti, si parla di *successo* se si verifica l'evento E e di *insuccesso* se E non si verifica (si verifica $\overline{E}$), senza che ciò prefiguri di per sé un evento piacevole.

È prassi far corrispondere al verificarsi di E il valore $(X = 1)$ e al verificarsi di $\overline{E}$ il valore $(X = 0)$, indicando con θ (o, spesso, anche con p) la probabilità che E si verifichi. In tal modo, si definisce una v.c. discreta che assume i soli due valori 0, 1 con probabilità, rispettivamente pari a:

$$Pr\left(\overline{E}\right) = Pr\left(X = 0\right) = 1 - \theta;\ \ Pr\left(E\right) = Pr\left(X = 1\right) = \theta\,.$$

Formalmente, una v.c. X discreta si definisce ***v.c. di Bernoulli*** se assume il valore 1 con probabilità θ ed il valore 0 con probabilità $1 - \theta$. Essa sarà indicata con $X \sim Ber(\theta)$ e la sua distribuzione di probabilità si può scrivere in modo compatto mediante la seguente espressione:

$$Pr\left(X = x\right) = \theta^x\,(1-\theta)^{1-x},\ x = 0, 1\,.$$

Infatti, tale espressione include sia il caso in cui si verifica E, cioè $(X = 1)$, con probabilità $\theta^1\,(1-\theta)^{1-1} = \theta$, che il caso in cui E non si verifica, cioè $(X = 0)$, con probabilità $\theta^0\,(1-\theta)^{1-0} = 1 - \theta$.

Schematicamente, quindi, la distribuzione di probabilità della v.c. di Bernoulli con parametro θ è la seguente:

Eventi	Valori di X	Probabilità
E	$(X=1)$	θ
$\overline{E}$	$(X=0)$	$1-\theta$
Totale	- - -	1

Tale v.c. è sempre ben definita se $0 \leq \theta \leq 1$. Inoltre, valore medio e varianza sono agevolmente deducibili:

$$\mathbb{E}(X) = \sum_{x=0}^{1} x\, P_r(X=x) = (0)\,(1-\theta) + (1)\,\theta = \theta;$$

$$Var(X) = \sum_{x=0}^{1} (x-\theta)^2\, P_r(X=x) = (0-\theta)^2\,(1-\theta) + (1-\theta)^2\,\theta = \theta(1-\theta).$$

o Quindi, per la v.c. $X \sim Ber(\theta)$ *il valore medio coincide con la probabilità dell'evento* E. Inoltre, la varianza è sempre compresa nell'intervallo $[0, 1/4]$, raggiungendo il suo massimo quando $\theta = 1/2$. Infatti, quando $\theta = 1/2$, la v.c. di Bernoulli deriva da una prova ove c'è la massima incertezza circa il verificarsi dell'evento E (come avviene, per esempio, nel lancio di una moneta ben equilibrata), essendo pari ad $1/2$ le probabilità sia di E che di $\overline{E}$.

Per contro, la varianza è nulla sia quando $\theta = 0$ (e quindi l'evento E è impossibile) che quando $\theta = 1$ (e quindi l'evento E è certo): in entrambi i casi, ovviamente, la v.c. di Bernoulli diventa una *v.c. degenere*, che concentra tutta la probabilità sull'evento $(X=0)$ nel primo caso e sull'evento $(X=1)$ nel secondo caso.

10.7 La variabile casuale Binomiale

Se si ripete, per n volte e nelle medesime condizioni, lo schema successo-insuccesso che, nel paragrafo precedente, ha condotto alla definizione della v.c. di Bernoulli, si genera una sequenza di n sottoprove indipendenti a ciascuna delle quali si può associare una v.c. di Bernoulli.

Tale schema equivale all'estrazione *con ripetizione* di n palline da un'urna che ne contiene H, di cui b bianche e $H-b$ nere, ove si ponga $\theta = b/H$ la probabilità (costante) di estrarre una pallina bianca in ciascuna estrazione.

In tali prove, interessa conoscere il numero X di volte in cui si è verificato l'evento E nel complesso delle n sottoprove, ripetute nelle medesime condizioni.

o Tale situazione è comune nei giochi, nei piani sperimentali, nel controllo di qualità, nelle indagini di laboratorio, nella teoria dei campioni. Tra gli esperimenti più usuali ove tale schema si presenta, citiamo i seguenti: nel lancio di n monete si è interessati al numero di «Testa» che complessivamente si verificano; nell'intervista casuale di n persone si vuole conoscere il numero di quelli favorevoli ad una certa proposta di legge; nell'ispezione di n barattoli scelti casualmente da una serie di confezioni interessa conoscere il numero di quelli che risultano difettosi; nella gestione quotidiana di n locomotori di una ferrovia locale interessa conoscere il numero di quelli che subiscono avarie o guasti; e così via.

o Formalmente, se una prova consiste nella ripetizione di n sottoprove indipendenti (per cui ciascuna sottoprova genera una v.c. di Bernoulli per la quale l'evento E =«successo», si verifica con probabilità θ costante), allora ogni possibile risultato della prova sarà la sequenza degli n risultati nelle n sottoprove. In altri termini, alcuni risultati saranno l'evento E ed altri l'evento $\overline{E}$. Pertanto, lo *spazio campione* della prova consisterà di tutte le possibili n-ple di sequenze di eventi E e $\overline{E}$.

Facciamo corrispondere ad ogni i-esima sottoprova, la v.c. $Y_i \sim Ber(\theta)$; quindi, in ciascuna sottoprova, la v.c. Y_i assumerà il valore 1 se si verifica E e il valore 0 se si verifica $\overline{E}$. Pertanto, il numero X di «successi» sul totale delle n sottoprove sarà esattamente pari al numero di volte che le v.c. Y_i hanno assunto il valore 1. Analogamente, gli insuccessi produrranno dei valori 0 nelle sottoprove corrispondenti.

Alla fine, per ottenere il numero X di successi sulle n sottoprove sarà sufficiente sommare i valori assunti dalle v.c. Y_i nelle singole sottoprove: tale somma $\sum_{i=1}^{n} Y_i$ è la determinazione della v.c. X che ora definiamo esplicitamente.

Una v.c. discreta X si definisce ***v.c. Binomiale*** se essa rappresenta il numero dei successi che si verificano in una sequenza di n sottoprove indipendenti nelle quali è costante la probabilità θ di un successo. Tale famiglia di v.c. dipende da due parametri (n, θ) e sarà indicata con: $X \sim Bin(n, \theta)$.

Poiché su n sottoprove, il numero di successi sarà $X = 0, 1, 2, \ldots, n$, la v.c. $X \sim Bin(n, \theta)$ assumerà i valori $0, 1, 2, \ldots, n$ in corrispondenza degli eventi corrispondenti come nello schema seguente:

non si è verificato alcun successo nelle n sottoprove	$(X=0)$
si è verificato esattamente 1 successo nelle n sottoprove	$(X=1)$
si sono verificati esattamente 2 successi nelle n sottoprove	$(X=2)$
.....	...
si sono verificati successi in tutte le n sottoprove	$(X=n)$

Per ottenere la distribuzione di probabilità di una v.c. $X \sim Bin(n, \theta)$, occorre calcolare la probabilità che sulle n sottoprove ve ne siano x, e solo x, che hanno dato luogo ad un successo.

o Una generica sequenza di x successi ed $(n-x)$ insuccessi ha probabilità $\theta^x(1-\theta)^{n-x}$ di verificarsi, stante la indipendenza delle singole sottoprove. Ebbene, tra le 2^n possibili sequenze che costituiscono gli eventi elementari di Ω, quelle che danno luogo ad esattamente x successi sono tutte e sole le sequenze nelle quali, combinandosi in ogni modo possibile, vi sono x successi ed $(n-x)$ insuccessi.

Il numero di tali sequenze è quello delle combinazioni di n elementi ad x ad x, cioè $\binom{n}{x}$ e ciascuna di esse avrà probabilità $\theta^x(1-\theta)^{n-x}$.

Poiché ognuna delle sequenze è incompatibile con le altre (perché le 2^n possibili sequenze sono eventi elementari e, quindi, costituiscono una partizione di Ω), la probabilità che X assuma il valore x, per $x = 0, 1, 2, \ldots, n$ è la somma di tali probabilità.

Pertanto, la distribuzione di probabilità della v.c. $X \sim Bin(n, \theta)$ è:

$$P_r(X=x) = \binom{n}{x} \theta^x\,(1-\theta)^{n-x}, \quad x = 0, 1, 2, \ldots, n.$$

Tale distribuzione è ben definita perché: $Pr(X = x) \geq 0$. Inoltre, la somma di tutte le probabilità è 1. Infatti, ricordando la formula del binomio di Newton, si ha:

$$\sum_{x=0}^{n} Pr(X = x) = \sum_{x=0}^{n} \binom{n}{x} \theta^x (1-\theta)^{n-x} = [\theta + (1-\theta)]^n = 1.$$

Quando $n = 1$, evidentemente, la v.c. Binomiale $X \sim Bin(1, \theta)$ coinciderà con la v.c. di Bernoulli $X \sim Ber(\theta)$.

La presenza dei coefficienti binomiali (e quindi dei fattoriali che li definiscono) può rendere lunga ed inefficiente la elaborazione esplicita di tali probabilità (anche quando n non è troppo elevato). Per questo, conviene calcolare solo la probabilità dell'evento $(X = 0)$ e, poi, tutte le altre in modo ricorsivo, grazie alle formule seguenti:

$$\begin{aligned} Pr(X = 0) &= (1-\theta)^n; \\ Pr(X = x+1) &= Pr(X = x) \frac{(n-x)\theta}{(x+1)(1-\theta)}, \quad x = 0, 1, \ldots, n-1. \end{aligned}$$

I momenti caratteristici della v.c. Binomiale risultano essere:

$$\mathbb{E}(X) = n\theta \; ; \qquad Var(X) = n\theta(1-\theta);$$

$$Asym(X) = \frac{1-2\theta}{\sqrt{n\theta(1-\theta)}} \; ; \qquad Kurt(X) = 3 + \frac{1-6\theta+6\theta^2}{n\theta(1-\theta)}.$$

∘ Il valore medio e la varianza possono essere derivati direttamente dalla distribuzione di probabilità della v.c. X, oppure, più semplicemente, ricordando che una v.c. $X \sim Bin(n, \theta)$ è la somma di n v.c. Y_i *indipendenti*, ciascuna delle quali è una v.c. di Bernoulli con identico parametrico pari a θ. Allora,

$$X = Y_1 + Y_2 + \cdots + Y_n.$$

Pertanto, utilizzando le proprietà del valore medio e della varianza per la somma di v.c. indipendenti (paragrafo 9.6), e ricordando che per una v.c. di Bernoulli il valore medio e la varianza valgono θ e $\theta(1-\theta)$, rispettivamente, si ha:

$$\begin{aligned} \mathbb{E}(X) &= \mathbb{E}(Y_1 + Y_2 + \cdots + Y_n) \\ &= \theta + \theta + \cdots + \theta = n\theta; \end{aligned}$$

$$\begin{aligned} Var(X) &= Var(Y_1 + Y_2 + \cdots + Y_n) \\ &= \theta(1-\theta) + \theta(1-\theta) + \cdots + \theta(1-\theta) = n\theta(1-\theta). \end{aligned}$$

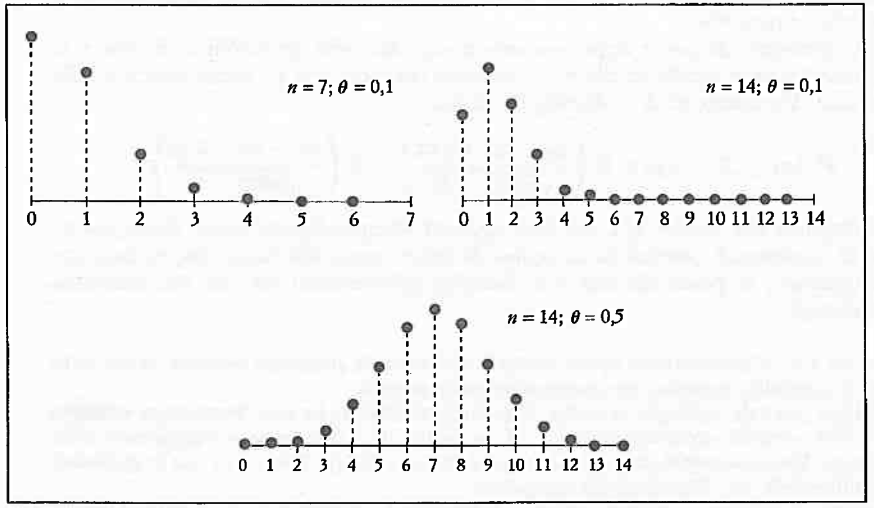

FIGURA 10.6. Distribuzioni di probabilità della v.c. Binomiale, al variare dei parametri n e θ.

Nella Figura 10.6 sono presentate, al variare dei parametri n e θ le distribuzioni di probabilità di alcune v.c. Binomiali.

Dai grafici di tali distribuzioni e dai momenti caratteristici si possono derivare alcune proprietà della v.c. Binomiale:

- La moda della v.c. binomiale è quell'unico intero M_0 compreso tra 0 e n tale che:

  $(n+1)\theta - 1 \leq M_0 \leq (n+1)\theta$.

  Se $(n+1)\theta - 1$ e $(n+1)\theta$ sono entrambi numeri interi, allora la distribuzione ha due mode distinte in corrispondenza di tali valori (e si dice che è *bimodale*). Per esempio, se $n = 14$, $\theta = 0.2$, allora la v.c. $X \sim Bin(14, 0.2)$ è bimodale, e possiede due mode in corrispondenza dei valori $x = 2$ e $x = 3$.

- la v.c. Binomiale è simmetrica solo quando $\theta = 1/2$ e per $n \to +\infty$; è asimmetrica positiva per $\theta < 1/2$ ed è asimmetrica negativa per $\theta > 1/2$;

- la somma di m v.c. *indipendenti* Binomiali tutte con uguale parametro θ è ancora una v.c. Binomiale. Formalmente, se $X_i \sim Bin(n_i, \theta)$, $i = 1, 2, \ldots, m$ sono v.c. indipendenti, allora: $\sum_{i=1}^{m} X_i \sim Bin(\sum_{i=1}^{m} n_i, \theta)$.

Una proprietà importante di tale v.c. (la cui giustificazione sarà fornita nel Capitolo seguente) afferma che al crescere di $n \to \infty$ la v.c. Binomiale tende alla v.c. Normale, e tale approssimazione è tanto più veloce (e quindi accurata) per

quanto più il parametro θ è vicino a 0.5 (quindi, quanto più la distribuzione è vicina alla simmetria).

Ciò consente di poter approssimare il calcolo delle probabilità di una v.c. Binomiale tramite quello di una v.c. Normale che possiede gli stessi valore medio e varianza. Pertanto, se $X \sim Bin(n, \theta)$, si ha:

$$Pr\left(x_1 \leq X \leq x_2\right) \simeq \Phi\left(\frac{x_2 - n\theta + 1/2}{\sqrt{n\theta(1-\theta)}}\right) - \Phi\left(\frac{x_1 - n\theta - 1/2}{\sqrt{n\theta(1-\theta)}}\right).$$

L'inserimento del valore 1/2 nei due estremi standardizzati viene detta *correzione di continuità*, perché ha lo scopo di tener conto del fatto che, in tale approssimazione, si passa da una v.c. discreta (Binomiale) ad una v.c. continua (Gaussiana).

- La v.c. Binomiale può essere interpretata come la *frequenza assoluta* di un certo evento E (definito *successo*) in n sottoprove indipendenti.

Allora, per tale tipologia di prove, si può anche definire la ***v.c. frequenza relativa*** $\mathcal{F}_n = X/n$ come il rapporto tra la v.c. $X \sim Bin(n, \theta)$ e il numero n complessivo delle sottoprove. Evidentemente, tale v.c. assume i valori $x = 0, 1/n, 2/n, \ldots, 1$ con le medesime probabilità della v.c. Binomiale da cui deriva.

Inoltre, è immediato dedurre che la v.c. frequenza relativa $\mathcal{F}_n$ così definita possiede valore medio e varianza espressi da:

$$\mathbb{E}(\mathcal{F}_n) = \frac{n\theta}{n} = \theta; \quad Var(\mathcal{F}_n) = \frac{n\theta(1-\theta)}{n^2} = \frac{\theta(1-\theta)}{n}.$$

Il risultato mostra che mentre valore medio e varianza crescono indefinitamente per la v.c. frequenza assoluta, per la v.c. frequenza relativa il valore medio è sempre θ mentre la varianza tende a 0 per $n \to \infty$. In altri termini, per $n \to \infty$, la v.c. frequenza relativa tende a diventare una *v.c. degenere* assumendo solo il valore θ con probabilità 1.

Sempre con riferimento a prove nelle quali interessa la dicotomia *successo-insuccesso*, numerose altre v.c. possono essere generate modificando le modalità di estrazioni o le caratteristiche di interesse della prova e delle relative sottoprove.

- Se le estrazioni sono effettuate *senza ripetizione* le probabilità si modificano ed il numero di successi in n sottoprove genera una nuova v.c. detta *v.c. Ipergeometrica*, di cui qui non ci occupiamo.
- Se in una prova bernoulliana, interessa il numero di sottoprove perché siano estratte almeno k palline bianche, allora tale numero è una realizzazione della *v.c. Binomiale negativa*; nel caso particolare di $k = 1$, tale v.c. si chiama *v.c. Geometrica*.
- Se, invece, la precedente prova è effettuata senza ripetizione ed interessa il numero di sottoprove necessarie affinché siano estratte almeno k palline bianche, allora tale numero è una realizzazione della *v.c. Ipergeometrica negativa*.

10.8 La variabile casuale di Poisson

A differenza delle precedenti, la v.c. di Poisson non è direttamente generata da una successione di prove bernoulliane (se non come processo limite), ma riguarda il numero di eventi registrati in un ambito circoscritto, temporale o spaziale.

- Tale v.c. fu derivata come limite della v.c. Binomiale da Poisson nel 1837 e, successivamente, nel 1898 Bortkiewicz la definì *v.c. degli eventi rari* applicandola al numero annuo di morti per calci di cavallo tra i soldati dell'armata prussiana, registrati ogni anno per 20 anni su 10 corpi di armata: l'adattamento tra la distribuzione empirica delle frequenze relative e quella teorica delle probabilità calcolate con la v.c. di Poisson era così notevole per i dati prescelti che fu all'origine dell'interesse per questo modello probabilistico.
- Successivamente, fu utilizzata da Student nel 1907 per il conteggio da particelle in gocce prelevate da un liquido e da Rutherford e Geiger nel 1910 per le variazioni dei raggi α emessi da una sostanza radioattiva (il polonio).
- Oggi, il suo uso si è generalizzato a tutti quei contesti nei quali occorre contare quanti eventi di un certo tipo si verificano in un intervallo temporale: quanti clienti arrivano ad uno sportello bancario ogni giorno, quanti incidenti si verificano in un certo percorso autostradale ogni settimana, quante chiamate arrivano ad un centralino telefonico ogni 10 minuti, quante auto sono in attesa di pagamento al casello autostradale ogni minuto, quanti studenti vengono al ricevimento ogni settimana, quanti giorni piovosi si verificano a Napoli ogni anno, e così via.
- I casi precedenti si riferiscono ad intervalli temporali ma rientrano nella stessa tipologia anche gli esempi seguenti: il numero di globuli rossi per ogni mm di sangue prelevato da un paziente, il numero di piccoli terremoti compresi tra il 3° e il 4° della scala Mercalli registrati settimanalmente nell'area vesuviana, il numero di incidenti mortali tra gli operai addetti ad un processo chimico pericoloso per ciascun impianto funzionante, il numero di errori commessi da un tipografo per ciascuna pagina di un libro, il numero di parole di tredici lettere per ciascuna pagina di un romanzo di un certo Autore, il numero di bombe cadute a Londra durante la II guerra mondiale per ciascun km^2 di superficie, e così via.
- Per semplificare la trattazione delle numerose tipologie di prove reali che originano situazioni ben adattate dalla v.c. di Poisson, si può indicare come «tempo» l'ambito specifico di riferimento, anche se quello che diremo si può agevolmente estendere ad altri contesti: territorio, classe sociale, pagina di libro, frase di un Autore, etc.

Definiamo *processo di Poisson* una successione di prove ripetute nelle medesime condizioni che soddisfano una serie di assiomi, che ora introdurremo.

Indichiamo con E l'evento prefissato, da studiare nell'intervallo di tempo $[0, T]$. Il modello probabilistico definito *processo di Poisson* è caratterizzato dalle tre condizioni seguenti:

- il verificarsi di E nell'intervallo (t_1, t_2) è indipendente dal verificarsi di E nell'intervallo (t_3, t_4), se tali intervalli non si sovrappongono;
- la probabilità del verificarsi di E in un intervallo infinitesimo $(t, t\,dt)$ è proporzionale ad un parametro $\theta > 0$ che caratterizza la prova;
- la probabilità che due eventi si verifichino nello stesso intervallo di tempo è un infinitesimo di ordine superiore rispetto alla probabilità che se ne verifichi uno solo.

Tali condizioni si possono formalizzare mediante equazioni differenziali nella funzione $p(x; \theta) = Pr\left(X = x\right)$ la cui unica soluzione, soggetta alla condizione di normalizzazione, definisce la distribuzione di probabilità per la v.c. X che ora esplicitiamo.

La ***v.c. di Poisson*** è il numero X di eventi E che si verificano in un definito intervallo temporale ed assume i valori: $x = 0, 1, 2, \ldots$ Quindi, X è una *v.c. discreta numerabile.*

La v.c. X di Poisson con parametro θ la indicheremo con $X \sim Po(\theta)$ e possiede la seguente distribuzione di probabilità:

$$Pr(X = x) = e^{-\theta} \frac{\theta^x}{x!}, \quad x = 0, 1, 2, \ldots$$

Essa è ben definita perché $Pr(X = x) > 0$ ed inoltre:

$$\sum_{x=0}^{\infty} Pr(X = x) = e^{-\theta} \sum_{x=0}^{\infty} \frac{\theta^x}{x!} = e^{-\theta} e^{\theta} = 1.$$

È conveniente ottenere tutte le probabilità mediante formule ricorsive, a partire da $Pr(X = 0)$. Infatti, si ha:

$$\begin{aligned} Pr(X = 0) &= e^{\theta}; \\ Pr(X = x+1) &= Pr(X = x) \frac{\theta}{(x+1)}, \quad x = 0, 1, 2, \ldots \end{aligned}$$

I momenti caratteristici della v.c. di Poisson risultano essere:

$$\mathbb{E}(X) = \theta; \quad Var(X) = \theta; \quad Asym(X) = \frac{1}{\sqrt{\theta}}; \quad Kurt(X) = 3 + \frac{1}{\theta}.$$

Nella Figura 10.7 sono presentate le distribuzioni di probabilità di alcune v.c. di Poisson, al variare del parametro θ.

Dai grafici di tali distribuzioni e dai momenti caratteristici si possono derivare alcune proprietà della v.c. di Poisson:

- il valore medio e la varianza della v.c. di Poisson coincidono entrambi con il parametro θ che caratterizza tale distribuzione di probabilità;
- la moda della v.c. di Poisson si verifica nel valore intero più vicino a θ; se il parametro θ è un numero intero, allora si presentano due mode di pari probabilità in corrispondenza di $x = \theta$ e di $x = \theta - 1$;
- la v.c. di Poisson è sempre asimmetrica positiva e tende a divenire simmetrica al crescere del parametro θ;
- per $\theta \to +\infty$ la distribuzione tende a divenire sempre più simmetrica e di forma campanulare (la curtosi tende a 3) e si può dimostrare che in tali casi essa è approssimata da una v.c. Normale;

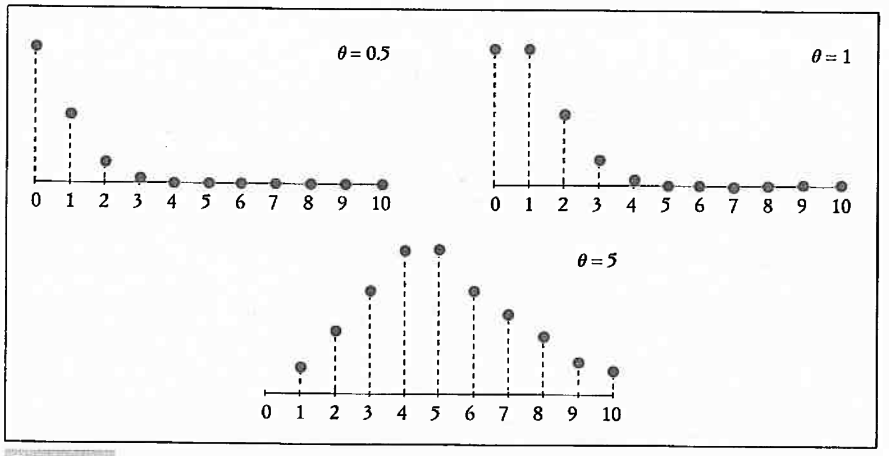

FIGURA 10.7. Distribuzione di probabilità della variabile casuale di Poisson, al variare del parametro θ.

- la somma di m v.c. *indipendenti* di Poisson con parametri rispettivamente pari a $\theta_1, \theta_2, \ldots, \theta_m$ è ancora una v.c. di Poisson che ha per parametro la somma di tali parametri; più precisamente, se $Y_i \sim Po(\theta_i)$, $i = 1, 2, \ldots, m$ sono v.c. indipendenti, allora: $\sum_{i=1}^{m} Y_i \sim Po(\sum_{i=1}^{m} \theta_i)$.

Infine, ricordiamo una proprietà della v.c. di Poisson che giustifica, in qualche modo, la sua originaria denominazione di «v.c. degli eventi rari».

- Data una v.c. Binomiale $X \sim Bin(n, \theta)$ con parametri n e θ, se $n \to +\infty$ e contemporaneamente $\theta \to 0$ (per cui è raro il verificarsi dell'evento E), in modo che comunque il prodotto $n\theta$ tenda ad una costante λ, cioè se $n\theta \to \lambda$, allora la distribuzione della v.c. Binomiale tende alla distribuzione della v.c. di Poisson con parametro λ.

Ambiente R

Principali modelli probabilistici

R10.1 Famiglie parametriche di variabili casuali

Rispetto alle più comuni famiglie parametriche (concernenti v.c. sia discrete che continue), molte delle quali non sono esaminate nel Capitolo 10 del testo, `R` offre un sistema integrato di risultati che è comune a tutte le v.c. A tal fine, schematizziamo per una generica v.v, che indichiamo con `varcas` (e che di volta in volta occore specificare), quali informazioni è possibile ottenere tramite `R` :

- `dvarcas(x,...)` genera la distribuzione di probabilità o la funzione di densità della v.c. nei valori x, con parametri specificati al posto dei ...
- `pvarcas(x,...)` genera la funzione di ripartizione della v.c. nei valori designati da x, con parametri specificati al posto dei ...
- `qvarcas(prob,...)` genera i quantili della v.c. nei valori designati da *prob*, con parametri specificati al posto dei ...
- `rvarcas(n,...)` genera n valori pseudo-casuali per la v.c., con parametri specificati al posto dei ...

Ovviamente, le funzioni che generano numeri pseudo-casuali non replicano gli stessi valori in simulazioni successive. Se si desidera conservare i valori simulati in un certo esperimento (per esempio, per ricontrollare i calcoli) occorre premettere alla generazione casuale la funzione `set.seed(numero)`, dove `numero` è un qualsiasi intero (da ricordare se si desiderano gli stessi valori): esso viene detto "seme" (=*seed*).

Sviluppiamo un esempio per la variabile casuale Uniforme continua sull'intervallo $[0, 1]$.

```
R==> runif(8);  ### Genera 8 numeri pseudo-casuali
[1] 0.13664 0.67605 0.04530 0.65272 0.99407 0.67764 0.91855
    0.46244

R==> runif(8);  ### idem, ma sono ovviamente diversi
[1] 0.57141 0.45193 0.16770 0.06826 0.89691 0.27383 0.93019
    0.54086

R==> set.seed(12345); runif(8); ### il seme è 12345
```

```
[1] 0.72090 0.87577 0.76098 0.88612 0.45648 0.16637 0.32510
    0.50922

R==> set.seed(12345); runif(8); ### stesso seme, stessi valori
[1] 0.72090 0.87577 0.76098 0.88612 0.45648 0.16637 0.32510
    0.50922

R==> set.seed(54321); runif(8); ### cambiano il seme ed i valori
[1] 0.42901 0.49843 0.17669 0.27439 0.21651 0.86636 0.04941
    0.20913
```

Nel seguito, con riferimento alle sole v.c. studiate nel testo, si specificheranno tali modelli (con le relative opzioni) mostrando anche come ottenere graficamente le rispettive distribuzioni di probabilità o funzioni di densità al variare dei parametri. In R si possono ottenere anche ulteriori opzioni rispetto a quelle qui elencate (per esempio, si possono ottenere i logaritmi delle probabilità, il complemento ad 1 della funzione di ripartizione, e così via).

R10.2 Variabile casuale Uniforme continua

Per un vettore x di valori ammissibili, la funzione `dunif(x,minimo, massimo)` fornisce la funzione di densità per i punti contenuti in x della v.c. Uniforme continua definita tra `minimo` e `massimo`; se tali valori non vengono esplicitati, si suppone che `minimo=0` e `massimo=1`.

Con le medesime opzioni, la funzione `punif(x,minimo, massimo)` fornisce la funzione di ripartizione calcolata nei punti contenuti in x. Invece, per un vettore di probabilità definito in `prob` la funzione `qunif(prob,minimo, massimo)` fornisce i quantili di tale v.c. calcolati per le probabilità contenuti in `prob`.

Infine, la funzione `runif(n,minimo, massimo)` genera n numeri pseudo-casuali da tale v.c. Così, l'istruzione `runif(100)` genera 100 numeri decimali pseudo-casuali tra 0 e 1.

Per esemplificare, mostriamo le istruzioni per ottenere la funzione di densità e la funzione di ripartizione di una v.c. uniforme sull'intervallo $(0,10)$; quindi, disegniamo $n = 1000$ punti sul quadrato unitario le cui ascisse ed ordinate sono determinate dalla generazione di valori pseudo-casuali della v.c. uniforme su $(0,1)$. Con l'occasione, si noti come per ottenere una rappresentazione più efficace (Figura R 10.1) si è utilizzata la funzione `layout()` per dividere lo spazio grafico in modo non standard.

```
R==> x=seq(0,10,by=0.01);
R==> densita=dunif(x,0,10);  frip=punif(x,0,10);
R==> assex=runif(1000);       assey=runif(1000);
# suddividi l'area grafica
R==> layout(matrix(c(1,2,3,3),2,2,byrow=TRUE));
### Funzione di densità
R==> plot(x,densita,type="l",lwd=2,xlab="f(x)",
    ylab="",ylim=c(0,0.12));
R==> title(main="Funzione di densità",cex=0.7);
### Funzione di ripartizione
R==> plot(x,frip,type="l",lwd=2,xlab="F(x)",ylab="");
R==> title(main="Funzione di ripartizione",cex=0.7);
### Punti pseudo-casuali nel quadrato unitario
R==> par(mar=c(3,4,2,2)+0.1);   ### amplia margini
R==> plot(assex,assey,xlab="",ylab="",pch=20);
```


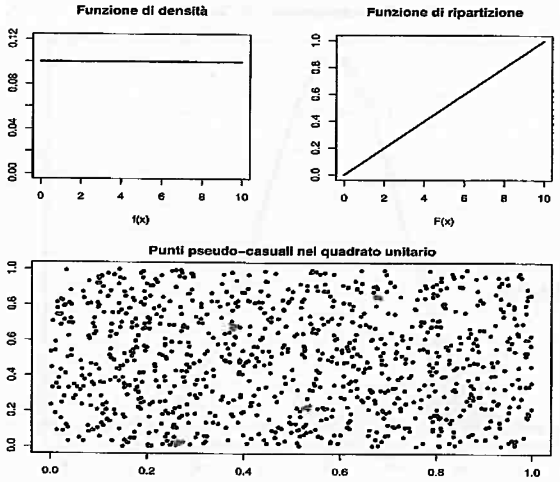

FIGURA R10.1. *Rappresentazioni connesse alla v.c. Uniforme continua.*

```
R==> title(main="Punti pseudo-casuali nel quadrato unitario",
    cex=0.6);
R==> par(mfrow=c(1,1));par(mar=c(5,4,4,2)+0.1) # ripristina
    opzioni
```

R10.3 Variabile casuale Normale (Gaussiana)

La v.c. Normale può essere utilizzata mediante le funzioni `dnorm()`, `pnorm()`, `qnorm()`, `rnorm()` con le opzioni `mean=media` e `sd=scarto quadratico medio` (attenzione! non la varianza). In assenza di opzioni, si suppone `mean=0, sd=1`, cioè ci si riferisce alla v.c. Normale standardizzata.

Nel seguito, mostriamo queste funzioni per la v.c. $Z \sim N(0,1)$. Le istruzioni seguenti (che generano la Figura R 10.2) evidenziano molteplici potenzialità grafiche di R (alcune qui presentate per la prima volta, come il riempimento di una porzione del piano) le quali dovrebbero essere agevoli da interpretare se riferite al grafico risultante.

```
R==> x=seq(-4,4,length=801);
R==> plot(x,dnorm(x),type="l",lwd=2.5,xaxp=c(-4,4,8),
       xlab="",ylab="");
R==> abline(h=0);
R==> title(main="Funzione di densità della v.c. N(0,1)");

R==> arrows(-1,0.01,0,0.01,code=3,length=0.1,lwd=1.5);
R==> text(-0.5,0.02,expression(sigma == 1));
R==> arrows(0,0,1,0.075,code=1,length=0.1,lwd=1.5);
R==> text(1.3,0.075,expression(mu == 0));

R==> dove=x[x<=-1.5];
```

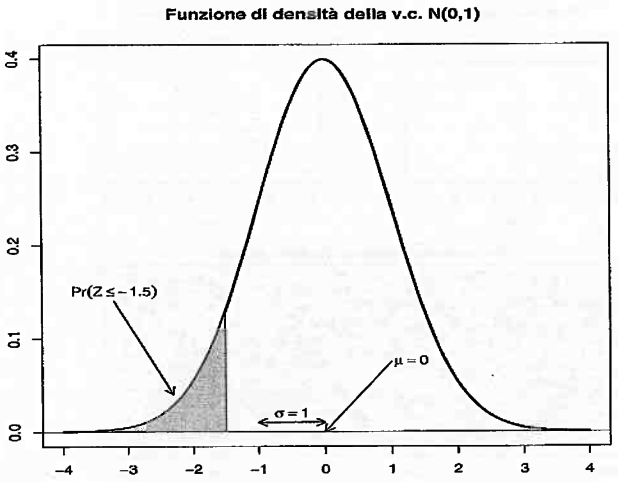

FIGURA R10.2. *Rappresentazioni connesse alla v.c. Normale standardizzata.*

```
R==> bordinf=c(-4,dove,dove[length(dove)],-4);
R==> bordsup=c(0,dnorm(dove),0,0);
R==> polygon(bordinf,bordsup,col=gray(0.7));
R==> text(-3.2,0.15,expression(Pr(Z<=-1.5)));
R==> arrows(-3.2,0.14,-2.3,0.04,code=2,length=0.1,lwd=1.5);
```

Per calcolare la probabilità dell'evento $(12 < X <= 16)$ per una v.c. $X \sim N(\mu, \sigma^2)$, con $\mu = 10$ e $\sigma^2 = 4$ (e, quindi, $\sigma = 2$) si possono utilizzare le tavole della v.c. Normale standardizzate (standardizzando gli estremi dell'intervallo, come indicato nel testo, Esempio 10.2) oppure sfruttare la funzione di ripartizione `pnorm(x,10,2)` mediante le istruzioni:

```
R==> pnorm(16,10,2)-pnorm(12,10,2)
[1] 0.15731
```

È possibile inserire come argomento anche un vettore di elementi; per esempio, per calcolare e disegnare il grafico della funzione di ripartizione della v.c. $X \sim N(10,4)$ (si ricordi sempre che $\sigma = 2$ designa lo scarto quadratico medio) basta definire il vettore dei valori in un conveniente intervallo attorno al valore medio e procedere con una sola istruzione al calcolo dell'intera funzione di ripartizione:

```
R==> x=seq(2,20,by=0.001);
R==> fripx=pnorm(x,10,2);
R==> p12=pnorm(12,10,2); p16=pnorm(16,10,2);
R==> plot(x,fripx,type="l",lwd=2,xaxp=c(0,20,10),
       xlab="",ylab="");

R==> title(main="Funzione di ripartizione della v.c. N(10,4)");
R==> abline(h=0); abline(v=0);
R==> segments(12,0,12,p12,lty="dotted");
R==> segments(0,p12,12,p12,lty="dotted");
```


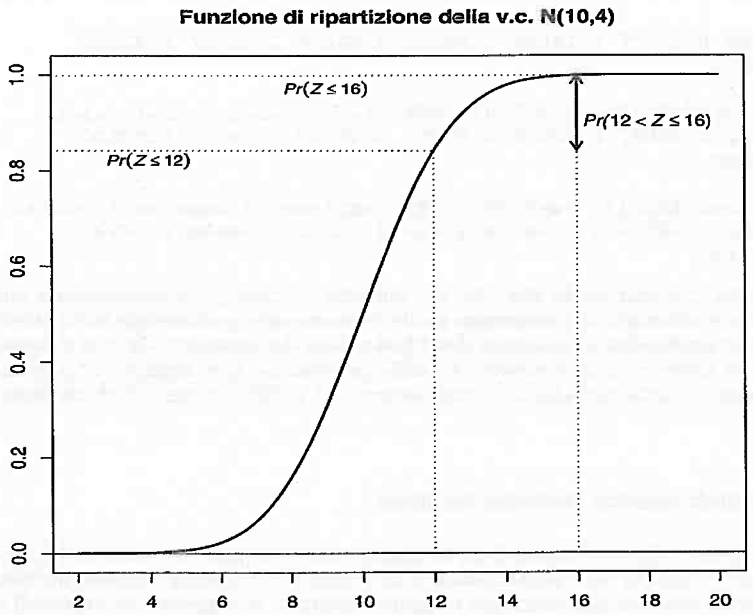

FIGURA R10.3. *Rappresentazioni connesse alla v.c. Normale standardizzata.*

```
R==> segments(16,0,16,p16,lty="dotted");
R==> segments(0,p16,16,p16,lty="dotted");
R==> arrows(16,p12,16,p16,code=3,length=0.1,lwd=1.5);
R==> text(9,0.97,expression(italic(Pr(Z <= 16))),cex=0.9);
R==> text(4,0.82,expression(italic(Pr(Z <= 12))),cex=0.9);
R==> testointerv=expression(paste(italic("Pr("),
       italic(12 < Z),italic(phantom()<= 16),")",sep=""));
R==> text(18,0.90,testointerv,cex=1.0);
```

La Figura R 10.3 mostra anche la relazione tra funzione di ripartizione e probabilità degli eventi ed è ottenuta sfruttando alcune delle funzioni ed opzioni grafiche di R .

Infine, con riferimento alla v.c. Normale standardizzata, è utile riassumere la corrispondenza tra le definizioni della teoria delle v.c., la simbologia del testo (utilizzata a livello internazionale) e le funzioni di R , come nel prospetto seguente. In tal modo, per esempio, è immediato tradurre l'espressione:

$$Pr\,(1 < Z \leq 3) = \Phi(3) - \Phi(1) = 0.99865 - 0.84134 = 0.15731$$

nella corrispondente istruzione in R :

```
R==> pnorm(3)-pnorm(1)
[1] 0.15731
```

TABELLA 10.1 *Funzioni in R per la v.c. Normale standardizzata* $Z \sim N(0,1)$.

Denominazione	*Simbologia*	*Definizione*	*Funzione di R*
Funzione di densità	$\phi(z)$	$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$	`dnorm(z)`
Funzione di ripartizione	$\Phi(z)$	$\int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$	`pnorm(z)`
Quantile (livello α)	z_α	$\Phi^{-1}(1-\alpha)$	`qnorm(α)`

R10.4 Variabili casuali Chi-quadrato, t ed F

Tali v.c. sono connesse alla v.c. Normale e vengono utilizzate per effettuare i più comuni test delle ipotesi statistiche (Cap.14). Pertanto, qui ci si limita a schematizzare il loro uso mediante la Tabella 10.2, utilizzando la sola funzione di densità (per le altre varianti basta cambiare la lettera iniziale della funzione) ed ignorando altre opzioni.

TABELLA 10.2 *Funzioni di densità e parametri di v.c. connesse alla v.c. Normale.*

Variabile casuale	*Funzione di densità*	*Parametri*
Chi-quadrato $\chi^2_{(g)}$	`dchisq(x,g)`	g =gradi di libertà
t di Student $t_{(g)}$	`dt(x,g)`	g =gradi di libertà
F di Fisher-Snedecor $F_{(g_1,g_2)}$	`df(x,g1,g2)`	g_1 = gradi di libertà del numeratore g_2 = gradi di libertà del denominatore

Per l'applicazione del test delle ipotesi, è utile calcolare i quantili di tali v.c. ai livelli di probabilità $\alpha = 0.05, 0.01, 0.001$ per prefissati gradi di libertà. Le istruzioni seguenti esemplificano questi casi con gradi di libertà posti pari a $g = 10$, $g = 10$, $g_1 = 3$, $g_2 = 8$, rispettivamente:

```
R==> alpha=c(0.05,0.01,0.001);
R==> chialpha=qchisq(alpha,10); chialpha;
[1] 3.9403 2.5582 1.4787
R==> talpha=abs(qt(alpha,10));  talpha;
[1] 1.8125 2.7638 4.1437
R==> falpha=qf(alpha,3,8);      falpha;
[1] 0.1130552 0.0363780 0.0076559
```

Stante la simmetria della v.c. t di Student, è conveniente utilizzare i quantili con il segno positivo (per questo, si è utilizzata la funzione `abs()`).

R10.5 Variabile casuale Uniforme discreta

Le istruzioni relative alla v.c. Uniforme discreta sono state già derivate nel paragrafo 9.1 nel caso particolare di $m = 6$. Essendo parametrizzate in funzione di qualsiasi $m > 1$, esse sono utilizzabili per qualsiasi v.c. Uniforme discreta definita sul supporto $\{1, 2, \ldots, m\}$.

R10.6 Variabile casuale Binomiale

La distribuzione di probabilità della v.c. $X \sim Bin(n, \theta)$ per i valori $x = 0, 1, \ldots, n$ è generata dalla funzione `dbinom(x,n,θ)`. Per esempio, le istruzioni:

```
R==> prob=dbinom(0:7,7,0.3);
R==> prob; sum(prob);
[1] 0.0823543 0.2470629 0.3176523 0.2268945 0.0972405 0.0250047
    0.0035721
[8] 0.0002187
[1] 1
R==> round(dbinom(0:7,7,0.3),digits=3);
[1] 0.082 0.247 0.318 0.227 0.097 0.025 0.004 0.000
```

generano la distribuzioni di probabilità della v.c. $X \sim Bin(10, 3)$ e verificano che la loro somma sia 1. L'ultima istruzione riproduce le stesse probabilità ma arrotondate con 3 decimali.

Le istruzioni successive consentono di disegnare sia la distribuzione di probabilità di tale v.c. che la corrispondente funzione di ripartizione di cui alla Figura R 10.4:

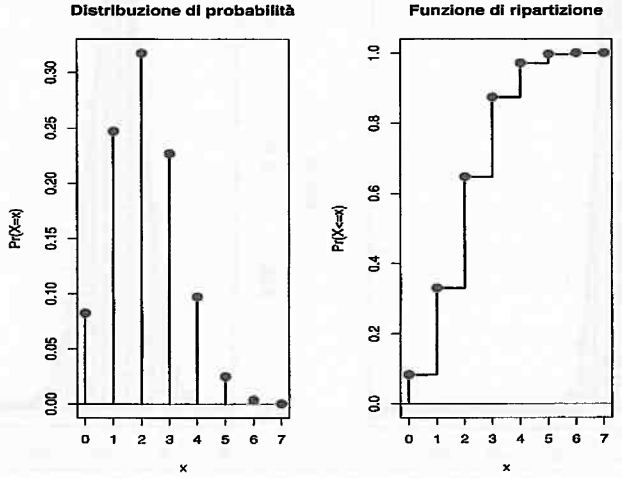

FIGURA R10.4. *Distribuzione di probabilità e funzioni di ripartizione della v.c.* $X \sim Bin(7, 0.3)$.

```
R==> n=7; theta=0.3;  x=0:n;
R==> probx=dbinom(x,n,theta);  fripx=pbinom(x,n,theta);
```

```
R==> par(mfrow=c(1,2))  # suddividi l'area in 2 regioni

### Distribuzione di probabilità
R==> plot(x,probx, type="h",lwd=2,
       xlab="x",ylab="Pr(X=x)");
R==> points(x,probx,pch=19,cex=1.5);
R==> abline(h=0);
R==> title(main="Distribuzione di probabilità",cex=0.8);
### Funzione di ripartizione
R==> plot(c(0,x),c(0,fripx), type="s",lwd=2, ylim=c(0,1),
       xlab="x", ylab="Pr(X<=x)");
R==> points(x,fripx,pch=19,cex=1.5);
R==> abline(h=0);
R==> title(main="Funzione di ripartizione",cex=0.8);
R==> par(mfrow=c(1,1))  # ripristina l'area del grafico
```

La Figura R 10.5 evidenzia l'approssimazione tra la distribuzione di probabilità della v.c. (discreta) Binomiale con parametri n e θ con la v.c. (continua) Normale con parametri $\mu = n\theta$ e $\sigma^2 = n\theta(1-\theta)$. Nello specifico, si confronteranno le v.c. $X \sim Bin(20, 0.05)$ – di valore medio $\mu_X = 20 \times 0.05 = 1$ e varianza $\sigma^2_X = 20 \times 0.05 \times 0.95 = 0.95$ – e le v.c. $Y \sim Bin(20, 0.45)$ – di valore medio $\mu_Y = 20 \times 0.45 = 9$ e varianza $\sigma^2_Y = 20 \times 0.45 \times 0.55 = 4.95$ – con le corrispondenti v.c. Normali di pari media e varianza.

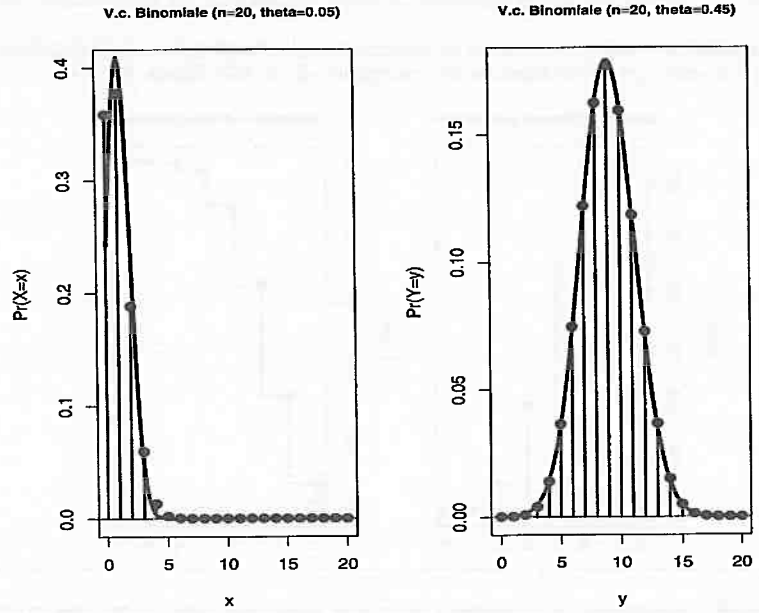

FIGURA R10.5. *Approssimazioni della v.c. Binomiale con la v.c. Normale.*

```
R==> n=20;thetax=0.05;
```



```
R==> x=0:n; probx=dbinom(x,n,thetax);
R==> mux=n*thetax; sigma2x=n*thetax*(1-thetax);
R==> xnorm=seq(0,20,length=100);
R==> densnormx=dnorm(xnorm,mux,sqrt(sigma2x));

R==> n=20;thetay=0.45;
R==> y=0:n; proby=dbinom(y,n,thetay);
R==> muy=n*thetay; sigma2y=n*thetay*(1-thetay);
R==> ynorm=seq(0,20,length=100);
R==> densnormy=dnorm(ynorm,muy,sqrt(sigma2y));

R==> par(mfrow=c(1,2))  # suddividi l'area del grafico

### Distribuzione di probabilità di X
R==> plot(x,probx, type="h",lwd=2,
       xlab="x",ylab="Pr(X=x)",ylim=c(0,0.4));
R==> points(x,probx,pch=19,cex=1.25);
R==> lines(xnorm,densnormx,lwd=3);
R==> abline(h=0);
R==> testox="V.c. Binomiale (n=20, theta=0.05)";
R==> title(main=testox,cex.main=0.8);

### Distribuzione di probabilità di Y
R==> plot(y,proby, type="h",lwd=2,
       xlab="y",ylab="Pr(Y=y)");
R==> points(y,proby,pch=19,cex=1.25);
R==> lines(ynorm,densnormy,lwd=3);
R==> abline(h=0);
R==> testoy="V.c. Binomiale (n=20, theta=0.45)";
R==> title(main=testoy,cex.main=0.8);
R==> par(mfrow=c(1,1))  # ripristina l'area del grafico
```

I grafici confermano che l'approssimazione migliora quando la distribuzione si avvicina ad una forma simmetrica (cioè, quando $\theta \simeq 0.5$).

R10.7 Variabile casuale di Bernoulli

La v.c. di Bernoulli è il caso particolare della v.c. Binomiale quando il parametro n (=numero delle prove) è uguale ad 1. Quindi, si possono utilizzare le istruzioni del paragrafo precedente ponendo $n = 1$.

D'altra parte, la v.c. di Bernoulli è una v.c. dicotomica che associa il valore 1 al cosiddetto "successo" ed il valore 0 all'"insuccesso"; per cui le funzioni di R possono essere utilizzate per generare successioni di numeri pseudo-casuali del tipo "successo/insuccesso" ove θ è la probabilità del "successo".

Per esempio, la sequenza di 100 risultati di lanci ripetuti di una moneta si ottengono con il comando `rbinom(100,1,0.5)`, cioè:

```
R==> rbinom(100,1,0.5)
  [1] 1 0 0 1 1 0 1 1 1 1 1 1 0 1 0 0 0 1 0 ................
 [38] ....................................................
 [75] 0 1 1 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 1 0 1 1 1 0 1 0
```

Per visualizzare come, al crescere del numero dei lanci, la frequenza relativa di un evento tenda progressivamente alla probabilità dell'evento, consideriamo il lancio di un

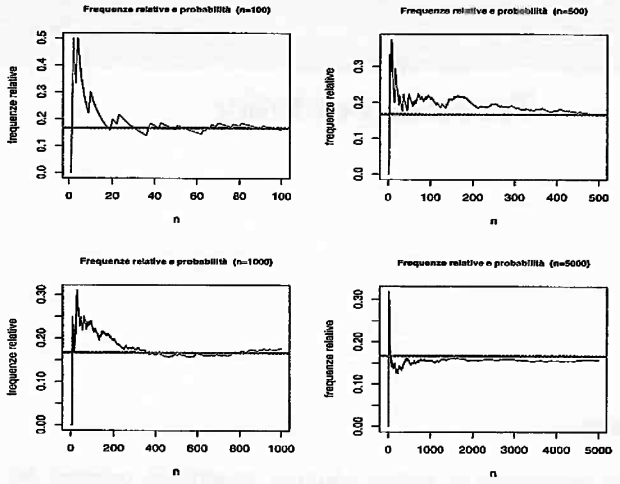

FIGURA R10.6. *Convergenza della frequenza relativa di un evento alla sua probabilità.*

dado e definiamo "successo" il verificarsi della faccia con 4 punti: tale evento ha probabilità $\theta = 1/6 \simeq 0.167$, per cui la generazione di numeri pseudo-casuali da una v.c. di Bernoulli con parametro $\theta = 1/6$ costituisce una simulazione al computer dell'esperimento fisico di lanci ripetuti ed indipendenti di un dado ben equilibrato. Naturalmente, per consentire al lettore di ripetere il medesimo esperimento qui raffigurato, occorre imporre al generatore di numeri pseudo-casuali di attivare esattamente le stesse operazioni (il che avviene in R premettendo l'istruzione `set.seed()`).

Le istruzioni seguenti effettuano tali lanci e calcolano sequenzialmente le frequenze relative dell'evento confrontandole con la probabilità costante di "successo". Il lettore noterà che, anziché ripetere per 4 volte le istruzioni per ciascun grafico, si è utilizzato un *loop* (cioè la funzione `for` che ripete un gruppo di istruzioni al variare dell'indice `j` in un certo insieme) per rendere efficiente la programmazione: questo ha richiesto di indicizzare il numero di simulazioni ed i titoli dei grafici.

```
R==> theta=1/6;
     titoli=c("Frequenze relative e probabilità  (n=100)",
              "Frequenze relative e probabilità  (n=500)",
              "Frequenze relative e probabilità  (n=1000)",
              "Frequenze relative e probabilità  (n=5000)");
R==> nsimul=c(100,500,1000,5000);
R==> set.seed(12345);   # per replicare lo stesso esperimento

R==> par(mfrow=c(2,2))  # suddividi l'area del grafico
R==> for(j in 1:4){
       conteggio=1:nsimul[j];
       vett=rbinom(nsimul[j],1,theta);
       cumfrel=cumsum(vett)/conteggio;
       plot(conteggio,cumfrel,type="l",lwd=0.8,col="blue",
            xlab="n",ylab="frequenze relative");
       title(main=titoli[j],cex.main=0.8);
       abline(h=theta,lwd=2,col="red");
}
R==> par(mfrow=c(1,1))  # ripristina l'area del grafico
```


La Figura R 10.6 (ottenuta con le istruzioni precedenti) mostra tali simulazioni per un numero crescente di esperimenti, e conferma l'assunto.

R10.8 Variabile casuale di Poisson

Le funzioni per la v.c. Poisson sono `dpois(x,theta)`, `ppois(x,theta)`, `qpois(x,theta)`, `rpois(x,theta)`, dove `theta` è il parametro che caratterizza tale v.c., cioè il valore medio indicato con θ nel testo.

La Figura R 10.7, che mostra la distribuzione di probabilità di tale v.c. al variare di θ, è ottenuta con le istruzioni seguenti; in esse, per evitare di replicare comandi simili, si è introdotto un *loop* che modifica di volta in volta i soli elementi differenti nei 4 grafici (generati per riga). Si osservi, in particolare, come l'opzione `ylim=c(0,1.2*max(vett))` incrementi verso l'alto l'asse delle probabilità di un 20% allo scopo di migliorare l'aspetto grafico delle rappresentazioni.

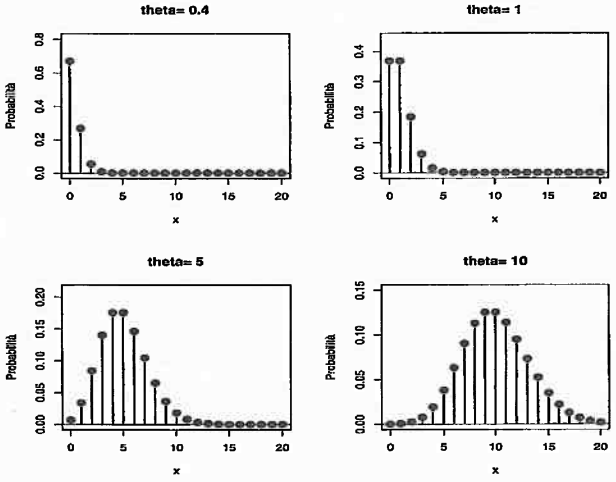

FIGURA R10.7. *Distribuzione di probabilità della v.c. di Poisson, al variare di θ.*

```
R==> theta=c(0.4,1,5,10);
R==> suppx=0:20;
R==> tit1=paste("theta=",theta[1]);  tit2=paste("theta=",theta[2]);
R==> tit3=paste("theta=",theta[3]);  tit4=paste("theta=",theta[4]);
R==> titoli=c(tit1,tit2,tit3,tit4);

R==> par(mfrow=c(2,2))  # suddividi l'area del grafico
     for(j in 1:4){
         vett=dpois(suppx,theta[j]);
```

```
        plot(suppx,vett,type="h",lwd=1.5,
           ylim=c(0,1.2*max(vett)),xlab="x",ylab="Probabilità");
        points(suppx,vett,pch=19,cex=1.5);
        title(main=titoli[j]);
        abline(h=0);
    }
R==> par(mfrow=c(1,1))  # ripristina l'area del grafico
```

La Figura R 10.7 conferma che la simmetria e la tendenza alla Normalità della v.c. di Poisson crescono con il parametro θ.

Capitolo undicesimo

Teoremi asintotici

11.1 Introduzione

I risultati che esporremo in questo capitolo riguardano teoremi del Calcolo delle probabilità la cui validità sussiste solo quando la numerosità delle v.c. in gioco aumenta. Per tali motivi sarebbe corretto premettere alla loro presentazione una discussione rigorosa sulle successioni di v.c. e sulle differenti definizioni di convergenza della successione verso una v.c. ben definita.

Queste problematiche, tuttavia, esulano dal livello di formalizzazione del nostro testo per cui l'affermazione «la successione di v.c. $X_n, n = 1, 2, \ldots$ *converge* alla v.c. Y» va interpretata, al crescere di $n \to +\infty$, in uno dei seguenti significati (che qui forniamo in modo intuitivo):

- ***convergenza in distribuzione***: la funzione di ripartizione di X_n (e, quindi, la distribuzione di probabilità se trattasi di v.c. discrete oppure la funzione di densità se trattasi di v.c. continue) tende per $n \to \infty$ ad approssimare la funzione di ripartizione di Y. In tali casi, scriveremo: $X_n \xrightarrow{d} Y$;

- ***convergenza in media quadratica***: il valore medio di $\mathbb{E}(X_n - Y)^2$ tende a 0 per $n \to \infty$.

Si comprende che, nel primo caso, la convergenza tra le v.c. viene interpretata tramite la corrispondente convergenza delle funzioni di ripartizione (che sono di natura matematica); nel secondo caso, invece, la convergenza viene limitata a quella di un particolare valore medio. Esistono altri modi di convergenza (anche più stringenti di quelli sopra enunciati), ma essi presuppongono formalizzazioni più complesse.

I teoremi che discuteremo qui di seguito hanno un ruolo fondamentale sia nell'Inferenza che nella modellistica per giustificare la validità di risultati approssimati; infatti, sono frequenti situazioni sperimentali nelle quali non è possibile accogliere tutte le assunzioni teoriche che sono alla base dei modelli probabilistici discusse nel capitolo precedente. Tra questi risultati un grande rilievo assumono le *leggi dei grandi numeri* e i *Teoremi Limite Centrale*.

Il capitolo si conclude con una presentazione essenziale della *teoria della simulazione* la quale, grazie al crescente sviluppo e alla massiccia diffusione degli strumenti di elaborazione automatica, rappresenta oggi un bagaglio essenziale della conoscenza statistica. In effetti, la simulazione rappresenta il primo legame tra la Teoria della probabilità (che con questo capitolo si conclude) e la Teoria dell'Inferenza statistica (che inizia dal capitolo successivo). Di conseguenza, il lettore troverà l'anticipazione di nozioni che andrebbero collocate nei capitoli successivi anche se, per ragioni di completezza, la loro posizione naturale è nell'ambito della Teoria della probabilità.

11.2 Leggi dei grandi numeri

Nella sua formulazione elementare, la legge dei grandi numeri deriva da alcune diffuse constatazioni sperimentali: ripetendo nelle medesime condizioni un esperimento casuale, al crescere del numero delle prove, la frequenza relativa di un evento tende a stabilizzarsi. Inoltre, per variabili quantitative rilevate in più sottoprove ripetute, la media rilevata su un campione di osservazioni si stabilizza al crescere della dimensione campionaria convergendo verso il valore medio della popolazione.

- Queste affermazioni generiche sono giustificate da numerosi e differenti teoremi che prendono il nome di *leggi dei grandi numeri* da quando Poisson utilizzò questo termine nel 1837. Nella versione originaria, esse risalgono a James Bernoulli che dimostrò uno di tali risultati nella «Ars Conjectandi», pubblicata postuma nel 1713.
- Per i motivi sopra accennati, i teoremi qui riportati riguardano la convergenza di una successione di v.c. ad una costante e le corrispondenti leggi dei grandi numeri vengono definite *deboli*.

Sia X_n una successione di v.c. generate da prove bernoulliane (cioè indipendenti) in ciascuna delle quali la probabilità che si verifichi l'evento E (il cosiddetto successo) è costante e vale θ. Quindi, ciascuna v.c. della successione è una v.c. $X_n \sim Ber(\theta)$, introdotta nel paragrafo 10.6.

Se ora definiamo la successione delle *v.c. somme parziali*:

$$S_1 = X_1;\ S_2 = X_1 + X_2;\ S_3 = X_1 + X_2 + X_3;\ \ldots;\ S_n = X_1 + X_2 + \cdots + X_n; \ldots$$

per quanto esplicitato nel paragrafo 10.7, ciascuna delle v.c. $S_1, S_2, \ldots$ è la somma di un numero finito di v.c. indipendenti di Bernoulli con lo stesso parametro θ: quindi, la v.c. S_n è una v.c. Binomiale con parametri n e θ, cioè: $S_n \sim Bin(n, \theta)$, per ogni $n = 1, 2, \ldots$.

Si ricordi che, poiché le v.c. di Bernoulli assumono valore 1 e 0 a seconda che l'evento E si verifichi oppure no nella singola sottoprova, la successione di v.c. S_n esprime la *frequenza assoluta* dei successi che si sono verificati nelle n prove bernoulliane. Di conseguenza, la v.c. $\mathcal{F}_n = S_n/n$ esprime la *frequenza relativa* di tali successi.

Per quanto già derivato alla fine del paragrafo 10.7, la v.c. $\mathcal{F}_n$ possiede valore medio $\mathbb{E}(\mathcal{F}_n) = \theta$ e varianza $Var(\mathcal{F}_n) = \theta(1-\theta)/n$ la quale, al crescere di n, tende a zero. Quindi, al crescere di n, *la v.c. frequenza relativa tende ad essere tutta accentrata sulla costante* θ, che coincide sia con il valore medio di tutte le v.c. della successione che con la probabilità costante dell'evento E.

Questo risultato (dovuto a Bernoulli), qui dimostrato per una particolare successione di prove indipendenti e somiglianti alla v.c. di Bernoulli, è stato successivamente esteso sia per v.c. non somiglianti che non indipendenti (ma con correlazione in qualche modo limitata). Il complesso di tali risultati teorici va sotto il nome di *leggi dei grandi numeri* e, sotto opportune condizioni (generalmente imponendo dei vincoli sulla variabilità delle v.c. coinvolte) permette di affermare che *la frequenza relativa di un evento converge alla sua probabilità.*

- Le leggi dei grandi numeri sono un sostegno importante alla impostazione frequentista del Calcolo delle probabilità ed hanno ampliato notevolmente le applicazioni della probabilità nei più svariati campi, grazie anche alla teoria della simulazione (paragrafo 11.5).
- In pratica, di fronte a qualsiasi prova ripetibile, è possibile fornire una valutazione accurata fin che si vuole della probabilità di uno specificato evento E replicando l'esperimento un numero grande di volte e registrando la frequenza relativa delle volte che tale evento si verifica: le leggi dei grandi numeri non solo dimostrano che tale frequenza si stabilizza con il crescere del numero delle prove (in effetti, la varianza della frequenza relativa tende a 0) ma anche che tale frequenza relativa converge verso una costante che è proprio la probabilità dell'evento E.
- L'ultima affermazione fa comprendere come tali leggi si applicano in numerosi campi (a volte, addirittura in modo inconscio, come se fosse ovvio il farlo). In ambiti differenti, si pensi a come viene stabilita la probabilità di guarigione o di efficacia di un farmaco; oppure la probabilità di attecchimento di una specie vegetale; oppure la probabilità di sopravvivenza di una cavia ad una prefissata dose di sostanza venefica; oppure la probabilità degli eventi vita/morte per le assicurazioni che sono determinati sulla base dell'analisi storica delle rispettive frequenze relative, e così via.

Tutte le discussioni sin qui svolte, e le connesse interpretazioni ed applicazioni, facevano riferimento alla frequenza relativa di un evento, ma le leggi dei grandi numeri sono state estese (da Khintchine) anche alla successione di v.c. indipendenti e somiglianti X_n tali che $\mathbb{E}(X_n) = \theta$, per ogni $n = 1, 2, \ldots$.

A tale riguardo, introduciamo la ***v.c. media aritmetica*** della sequenza $(X_1, X_2, \ldots, X_n)$, che è definita da:

$$\frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Allora, se la successione di v.c. indipendenti e somiglianti X_n è tale che $\mathbb{E}(X_n) = \theta$, per ogni $n = 1, 2, \ldots$, le leggi dei grandi numeri affermano che le medie aritmetiche delle v.c. X_n tendono al valore medio (costante) delle v.c. X_n.

Più brevemente, sotto le condizioni indicate le v.c. medie aritmetiche $\sum_{i=1}^{n} X_i/n$ tendono, al crescere di n, alla quantità costante $\mathbb{E}(X_n) = \theta$.

o Il fatto che la successione delle medie aritmetiche di v.c. tenda ad un valore medio costante ha importanti implicazioni per l'Inferenza statistica. Ma essa è un risultato scientifico notevole anche in altri campi; per esempio, la dimostrazione del teorema di Weierstrass (uno dei risultati più importanti dell'Analisi matematica) è resa semplice grazie alle leggi dei grandi numeri.

11.3 Teorema Limite Centrale

Il Teorema Limite Centrale è uno dei risultati più importanti e più utili della Teoria della probabilità e della Statistica e possiede, anche a prescindere dalle numerose applicazioni, un valore scientifico rilevante per la teoria della conoscenza, l'evoluzione naturale e la logica delle decisioni umane.

o Il Teorema Limite Centrale (come il teorema di Pitagora o la gravitazione universale) è una delle poche acquisizioni scientifiche per le quali vi è stata una evidenza empirica antica e costante che, solo ad un certo punto, viene formalizzata in modo rigoroso. Anche se la sua esplicitazione giunge piuttosto in ritardo, esso era ben evidente nelle misurazioni ripetute condotte da Galileo, nella distribuzione delle caratteristiche fisiche delle specie viventi, dal grafico dei coefficienti binomiali esposti nel triangolo di Pascal e noti ai cinesi sin dall'antichità, nella forma dei vulcani, nelle strategie istintive e comuni che suggeriscono un posizionamento centrale nella gran parte delle scelte umane, e così via.

o Trattandosi di un risultato matematico che ha un notevole impatto applicativo, esso non vale in forma esatta in tutte le situazioni reali. Ciò nonostante, esso resta un formidabile strumento intellettuale esattamente come i teoremi della Geometria: nessuno può evitare di utilizzare triangoli rettangoli, anche se nessuno riuscirà a costruirne uno solo che perfettamente rispetti il teorema di Pitagora. Così, non si può sottovalutare il Teorema Limite Centrale solo perché è impossibile trovare un fenomeno reale che sia perfettamente Normale.

o Il Teorema Limite Centrale è stato definito «centrale» (da Polya nel 1920) perché v.c. di forma qualunque tendono a convergere verso una distribuzione «centrata» sulla media, che è la v.c. Normale.

o Nella sua formulazione più ingenua, il Teorema Limite Centrale asserisce che la somma di una successione di v.c. indipendenti ed identicamente distribuite (purché di varianza finita) converge in distribuzione alla v.c. Normale. Tale risultato è stato esteso e generalizzato mediante risultati sempre più specifici, ma anche complessi nella loro formulazione analitica; nel seguito, ci limiteremo a presentare il Teorema Limite Centrale nei due casi più semplici e comuni.

o La prima versione del Teorema Limite Centrale come somma di v.c. di Bernoulli fu dedotta, non rigorosamente, da De Moivre nel 1718 e poi nel 1733 nel caso di $\theta = 1/2$. Essa fu ampiamente utilizzata da Gauss nel 1809 per derivare la sua teoria degli errori accidentali; ma la dimostrazione e la formalizzazione di tali risultati si devono all'opera di Laplace, a partire dal 1812 (infatti, il teorema di De Moivre è, quasi sempre, definito di De Moivre e Laplace).

Abbiamo già anticipato che esistono numerose versioni del Teorema Limite Centrale che cercano di ampliare le sue possibilità di applicazione rimuovendo progressivamente le restrizioni che ne sono alla base. Per esempio, il Teorema Limite Centrale è stato esteso anche alla somma standardizzata di successioni di v.c. non indipendenti (ma con correlazione che tende a decrescere secondo leggi prefissate) ovvero alla somma di un numero aleatorio di v.c.

Tra le condizioni sulle v.c. addendi (che accelerano la convergenza alla Normalità della successione di somme standardizzate di v.c.), possiamo elencare il seguente ordine di priorità:

i) *unimodalità* delle distribuzioni;

ii) *simmetria* delle distribuzioni;

iii) *somiglianza* delle distribuzioni;

iv) *indipendenza* delle v.c.

Quindi, la somma di v.c. unimodali, simmetriche, identicamente e indipendentemente distribuite conduce nel modo più veloce possibile alla convergenza della v.c. risultante verso la distribuzione Normale. Al contrario, la somma di v.c. plurimodali, asimmetriche, differenti tra loro e non indipendenti ritarda al massimo tale convergenza.

Qui premettiamo solo una elencazione delle versioni più note dei Teoremi Limite Centrale, esplicitando poi la formulazione solo per i primi due.

- Il *teorema di De Moivre e Laplace* dimostra la tendenza alla Normalità della successione di somme standardizzate di v.c. indipendenti e somiglianti del tipo Bernoulli per cui, alla fine, si risolve nella convergenza di una v.c. Binomiale standardizzata verso la v.c. Normale standardizzata.

- Il *teorema di Lindeberg e Lévy* generalizza quello di De Moivre e Laplace esaminando la convergenza di somme di v.c. indipendenti e identicamente distribuite con varianza finita. In effetti, esso vale anche senza vincoli sulla varianza purché si ammetta un comportamento sufficientemente regolare nelle code della distribuzione della successione di v.c.

- Il *teorema di Lyapunov* esplicita condizioni sufficienti affinché la convergenza alla v.c. Normale avvenga per somme di v.c. indipendenti (ma non somiglianti) ponendo condizioni sull'esistenza dei momenti assoluti del terzo ordine.

- Il *teorema di Lindeberg e Feller* evidenzia condizioni necessarie e sufficienti per la convergenza alla v.c. Normale di somme di v.c. indipendenti (non somiglianti) connesse al peso relativo tra le varianze delle v.c. addendi.

- Il *teorema di Lévy* deriva da quello di Lindeberg e Feller nel caso di v.c. indipendenti, somiglianti e con varianza finita.

- Il *corollario di Fisz* estende il teorema di Lindeberg e Feller a v.c. definite su un supporto finito.

Nella sua forma originaria, data una successione di v.c. indipendenti e somiglianti $\{X_i,\ i = 1, 2, \dots\}$, con valore medio $\mathbb{E}(X_i) = \mu$ e varianza $Var(X_i) = \sigma^2 < +\infty$, il Teorema Limite Centrale riguarda la convergenza della v.c. somma $S_n = \sum_{i=1}^{n} X_i$ quando $n \to +\infty$.

Poiché tale v.c. avrà sia il valore medio ($= n\mu$) che la varianza ($= n\sigma^2$) che, al crescere di n, tenderanno entrambi a diventare infiniti, occorre considerare la ***v.c. somma standardizzata*** Z_n che è definita da:

$$Z_n = \frac{S_n - \mathbb{E}(S_n)}{\sqrt{Var(S_n)}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu}{\sigma/\sqrt{n}}.$$

Per la v.c. Z_n, come è noto, il valore medio è sempre 0 e la varianza è sempre 1.

Nella precedente definizione, l'ultimo passaggio mostra come, alla fine, la convergenza della v.c. somma S_n sia la stessa della *v.c. media aritmetica* $\frac{1}{n}\sum_{i=1}^{n} X_i$. Infatti, trattandosi di una v.c. standardizzata, la divisione della v.c. originaria per n non modifica il risultato.

■ ***Teorema Limite Centrale di De Moivre e Laplace***

Se X_n è una successione di v.c. indipendenti e identicamente distribuite come una v.c. di Bernoulli con parametro $\mathbb{E}(X_n) = Pr(X = 1) = \theta$, allora la v.c. somma standardizzata Z_n tende ad avere la stessa distribuzione della v.c. Normale standardizzata $Z \sim N(0, 1)$.

Scriveremo:

$$Z_n = \frac{S_n - n\theta}{\sqrt{n\theta(1-\theta)}} \xrightarrow{d} Z \sim N(0, 1).$$

■ ***Teorema Limite Centrale di Lindeberg e Lévy***

Se X_n è una successione di v.c. indipendenti e identicamente distribuite con valore medio μ e varianza $0 < \sigma^2 < +\infty$, allora la v.c. somma standardizzata Z_n tende ad avere la stessa distribuzione della v.c. Normale standardizzata $Z \sim N(0, 1)$.

Scriveremo:

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z \sim N(0, 1).$$

○ Nei Capitoli precedenti, il Teorema Limite Centrale è stato già applicato (con gli opportuni accorgimenti dovuti alla natura eventualmente discreta e/o asimmetrica delle distribuzioni), quando si è parlato di «tendenza» o di approssimazione della distribuzione di una v.c. alla v.c. Normale. Tale convergenza alla v.c. Normale vale per la somma di v.c. Uniformi continue (paragrafo 10.2) e discrete (paragrafo 10.5), per la somma di v.c. di Bernoulli (e quindi della v.c. Binomiale: paragrafo 10.7) e di v.c. di Poisson (paragrafo 10.8), nonché per le v.c. Chi-quadrato e di Student (paragrafo 10.4), al crescere dei rispettivi gradi di libertà.

○ Tuttavia, l'importanza del Teorema Limite Centrale deriva da considerazioni più generali. Il fatto che la v.c. Z_n, *somma standardizzata di v.c. di qualunque tipo* (purché indipendenti e con vincoli che servono solo ad escludere situazioni estreme), converga ad una distribuzione Normale standardizzata, a prescindere dai modelli probabilistici che generano le singole v.c., implica che *tutte le volte che un fenomeno reale può essere interpretato come la somma, oppure la media, di un gran numero di cause indipendenti, nessuna delle quali ha prevalenza sulle altre, è ragionevole attendersi che la distribuzione di probabilità di quel fenomeno sia approssimabile mediante la distribuzione della v.c. Normale.*

Tale affermazione continua a valere se il fenomeno è la combinazione lineare di un gran numero di cause, perché ciò non ha effetto sulla standardizzazione delle v.c. somma o media aritmetica.

○ L'altezza di una persona adulta è il risultato della combinazione di un gran numero di cause tra cui l'altezza dei genitori e dei nonni (paterni e materni), l'alimentazione ricevuta durante la gestazione e la prima infanzia, eventuali traumi, l'abitudine a pratiche sportive, la latitudine in cui si è nati, e così via. Queste cause si combinano assieme e generano quella che è l'altezza di una persona facendo presumere che la v.c. «risultante» sia ben approssimata da una v.c. Normale. I dati raccolti e più volte esaminati nel data set STUDENTI mostrano che effettivamente l'altezza è ben approssimata da una v.c. Normale.

○ Il punteggio che si riceve sottoponendosi ad una serie numerosa di test a risposta multipla è la risultante di numerose cause tra le quali: la preparazione generale del soggetto, la conoscenza degli argomenti specifici selezionati dall'esaminatore, l'attitudine verso la materia, i fattori psicologici e momentanei, e così via: anche se difficili da quantificare, questi effetti determinano con pesi differenti l'esito per cui il punteggio finale si distribuisce approssimativamente come una v.c. Normale.

○ Le caratteristiche fisiche di una foglia sono il risultato di una complessa evoluzione biologica nella quale intervengono numerosi fattori indipendenti: acqua, condizioni meteorologiche, genetiche ed ambientali, danni accidentali, competizione con altri vegetali, esposizione solare, etc. Questi, con pesi differenti, contribuiscono a determinare le caratteristiche in oggetto. Quindi, la lunghezza e le altre variabili della stessa natura (larghezza, spessore, peso della foglia) possono legittimamente essere approssimate mediante la v.c. Normale, come peraltro appare confermato dalle osservazioni raccolte nel data set CAMELLIA.

○ La misura della luminosità di una stella, il peso dei frutti di una specie vegetale, la rugosità di una superficie metallica, i tempi di percorrenza di un percorso urbano nell'ora di punta, le variazioni dell'inflazione mensile in un determinato Paese, etc. costituiscono ulteriori esempi per i quali è verosimile che il Teorema Limite Centrale produca una approssimazione soddisfacente alla v.c. Normale di tali quantità. Infatti, in tutti questi casi, il fenomeno in oggetto può essere concepito come la determinazione di una v.c. che risulta dalla somma di un gran numero di v.c. indipendenti e ciascuna con variabilità finita.

○ Altri fenomeni, come per esempio la durata delle conversazioni telefoniche, i tempi di diffusione delle informazioni e la distribuzione personale del reddito, derivano dall'interazione moltiplicativa tra le varie cause che presiedono alla loro generazione. Quindi, è ragionevole attendersi che sia il logaritmo del fenomeno (o qualche altra trasformazione) ad essere distribuito come una v.c. Normale, come peraltro le osservazioni reali confermano.

o Una dimostrazione *meccanica* del Teorema Limite Centrale è stata proposta da Galton nel 1889 nel suo libro «Natural Inheritance» grazie ad uno strumento (da lui denominato «*quincunx*») nel quale una pallina entra in un imbuto e, costretta dalla gravità ad un percorso verticale, è ostacolata nella sua caduta, da un numero n di chiodi che la deviano a sinistra o a destra con pari probabilità, sino a raccogliersi in un certo numero di caselle equidistanti. Al crescere del numero n di ostacoli interposti, la distribuzione di frequenza osservata di un numero elevato di palline che si fanno cadere converge a quella di una v.c. Normale. Sono evidenti le variazioni da apporre agli ostacoli per generare effetti differenti dalla indipendenza, dalla simmetria e dalla somiglianza delle singole distribuzioni e studiare empiricamente le distribuzioni risultanti da tali modifiche. Anziché costruire uno strumento simile (che è una sorta di flipper con ostacoli equi-distribuiti, visibile spesso nei Musei della scienza e della tecnica), ciascuno può simulare «la pallina che cade» e visualizzare la risultante distribuzione di frequenza, mediante qualsiasi software che possieda un generatore di numeri pseudo-casuali uniformi (paragrafo 11.5). Oppure, molto più semplicemente, basta lanciare 6-8 dadi ripetutamente e riportare la distribuzione della «somma dei punti risultanti sulle facce dei dadi»: dopo qualche centinaio di lanci la distribuzione risultante sarà assai vicina a quella di una v.c. Normale.

Un posto di rilievo, anche per ragioni storiche, merita la cosiddetta ***legge degli errori accidentali*** (formulata da vari Autori e autorevolmente dimostrata da Gauss) secondo la quale in una serie ripetuta e indipendente di misurazioni di una costante fisica, il cui valore vero è μ, si genera una v.c. scarto $X - \mu$ rispetto al valore vero μ che, al crescere delle prove, tende a distribuirsi come una v.c. $N(0, \sigma^2)$, dove σ^2 è una quantità che è funzione inversa della *precisione* dello strumento di misura utilizzato.

In tale situazione sperimentale, infatti, sono presenti le condizioni tipiche per l'applicabilità dei Teoremi Limite Centrale: stanchezza fisica dell'osservatore, abilità ed esperienza, imprecisione nella lettura, temperatura, umidità atmosferica, etc., cioè una successione piuttosto grande di v.c. indipendenti di varianza finita la cui combinazione lineare (che determina la misurazione effettiva) si adatta molto bene ad una v.c. Normale.

La legge degli errori accidentali è alla base di qualsiasi misura fisica, chimica, biologica ed elettronica e spiega perché è così diffusa l'abitudine di accompagnare la cosiddetta «misura vera» con un intervallo di $\pm 2\sigma$: infatti, come abbiamo visto nel paragrafo 10.3, vi è probabilità superiore al 95% che i valori della v.c. Normale ricadano in tale intervallo.

o Da questi importanti risultati (che sfrutteremo ampiamente nei successivi capitoli), non bisogna trarre l'erronea conclusione che la somma di v.c. tende sempre alla v.c. Normale ovvero che tutti i fenomeni del mondo reale siano ben approssimati da una v.c. Normale. Esistono situazioni nelle quali la convergenza avviene verso altre v.c. oppure non avviene alcuna forma di convergenza, oppure la v.c. è radicalmente differente da una distribuzione unimodale e campanulare, come molte delle v.c. definite nel Capitolo 10 confermano.

Resta indubbio, però, che la v.c. Normale è un punto di riferimento anche per v.c. che si allontanano da tale distribuzione; a tale riguardo, spesso, si cerca di trasformare il fenomeno che si osserva in modo da migliorare la sua rappresentazione in modo che sia, almeno approssimativamente, assimilabile a quello generato da una v.c. Normale. Ciò è giustificato dal fatto che per v.c. Normali i metodi inferenziali della Statistica (trattati nella parte terza di questo volume) sono semplici, immediati e particolarmente efficaci.

Presentiamo, ora, un esempio nel quale si mostra come, al crescere del numero degli addendi, sia confermata la validità del teorema di Lindeberg e Lévy per v.c. indipendenti e somiglianti.

Esempio 11.1

Per fornire una conferma del Teorema Limite Centrale ed una visualizzazione della tendenza della distribuzione della v.c. somma a quella della v.c. Normale, consideriamo una successione di v.c. indipendenti e tutte Uniformi continue (paragrafo 10.2) definite sull'intervallo $[0,1]$, cioè $X_i \sim U(0,1)$, $\forall i = 1, 2, \ldots$ Quindi, ciascuna v.c. X_i possiede valore medio $\mathbb{E}(X_i) = 1/2$ e varianza $Var(X_i) = 1/12$.

Al variare di $n = 1, 2, \ldots$, la v.c. somma $S_n = \sum_{i=1}^{n} X_i$ possiede valore medio e varianza espressi, rispettivamente, da:

$$\begin{aligned} \mathbb{E}(S_n) &= \sum_{i=1}^{n} \mathbb{E}(X_i) = \frac{n}{2}; \\ Var(S_n) &= \sum_{i=1}^{n} Var(X_i) = \frac{n}{12}; \end{aligned}$$

per cui per la v.c. somma standardizzata Z_n, grazie al Teorema Limite Centrale di Lindeberg e Lévy, converge alla v.c. Normale standardizzata.

Formalmente, si avrà:

$$Z_n = \frac{S_n - n/2}{\sqrt{n/12}} = \sqrt{\frac{3}{n}} \left(2 \sum_{i=1}^{n} X_i - n \right) \xrightarrow{d} Z \sim N(0,1) .$$

Ora, è possibile esaminare come si comporta la distribuzione della v.c. S_n per $n = 1, 2, 3, \ldots$ e, al crescere di n, controllare la bontà dell'approssimazione di tali funzioni di densità con quella di una v.c. Normale che ha lo stesso valore medio e la stessa varianza della v.c. S_n.

Ciò è relativamente agevole grazie ad un risultato (ottenuto da Laplace, nel 1814) secondo il quale la funzione di densità della v.c. S_n sopra definita è uguale, *in modo esatto*, alla seguente espressione:

$$f_{S_n}(s) = \frac{1}{(n-1)!} \sum_{r=0}^{k} (-1)^r \binom{n}{r} (s-r)^{n-1}, \quad k \le s \le k+1, \quad k = 0, 1, 2, \ldots, n-1.$$

La precedente formula mostra che la funzione di densità di una somma di v.c. Uniformi continue è una costante se $n = 1$; è la giustapposizione di due rette se $n = 2$; è la giustapposizione di tre archi di parabola se $n = 3$, etc.

In generale, la somma di n v.c. Uniformi continue ha una funzione di densità che si ottiene mediante la giustapposizione di n polinomi di grado $n - 1$, ciascuno dei quali è definito su sotto-intervalli di eguale ampiezza sul supporto della v.c. somma.

Per esempio, quando $n = 4$, la v.c. S_4, somma di 4 v.c. indipendenti e Uniformi continue sul supporto $[0, 1]$, ha per funzione di densità la seguente espressione:

$$f_{S_4}(s) = \begin{cases} \frac{1}{6}s^3, & 0 \le s \le 1; \\ \frac{1}{6}(-3s^3 + 12s^2 - 12s + 4), & 1 \le s \le 2; \\ \frac{1}{6}(3s^3 - 24s^2 + 60s - 44), & 2 \le s \le 3; \\ \frac{1}{6}[-(s-4)^3)], & 3 \le s \le 4. \end{cases}$$

Come si vede, tale funzione di densità è ottenuta dalla giustapposizione di 4 polinomi di 3° grado definiti su 4 sotto-intervalli di uguale ampiezza del supporto $[0, 4]$.

Ebbene, la Figura 11.1 rappresenta le funzioni di densità delle v.c. S_n, per $n = 1, 2, 3, 4$. È molto evidente che la forma della funzione di densità della v.c. somma tende velocemente a quella di una v.c. Normale che ha lo stesso valore medio e la stessa varianza della v.c. somma. Si vede chiaramente che tale approssimazione migliora al crescere di n e può già ritenersi soddisfacente per $n = 4$.

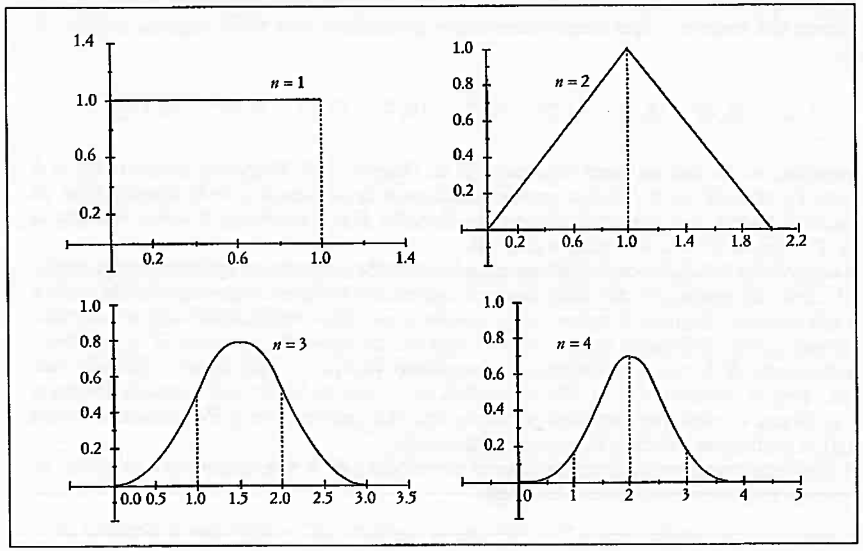

FIGURA 11.1. Funzioni di densit della variabile casuale Somma di n variabili casuali Uniformi sull'intervallo $(0, 1)$, per $n = 1, 2, 3, 4$.

□

11.4 Altri teoremi asintotici

La Teoria della probabilità include numerosi altri teoremi asintotici che completano, secondo prospettive differenti, sia la legge dei grandi numeri che il Teorema Limite Centrale, e molti di essi sono utilizzati nell'ambito dei metodi inferenziali. Qui ci limitiamo ad accennare al solo ***teorema di Glivenko e Cantelli*** la cui importanza è espressa dal fatto che esso è anche chiamato il *teorema fondamentale della Statistica.*

Nella sua essenza, il teorema afferma che la funzione di ripartizione empirica $F_n(x)$ converge *quasi certamente* e uniformemente alla funzione di ripartizione teorica $F(x)$. Questa forma di convergenza è molto forte perché afferma che è un evento certo (cioè ha probabilità pari a 1) un determinato comportamento asintotico.

Ora, la funzione di ripartizione empirica $F_n(x)$ è una funzione deducibile dalle osservazioni mentre $F(x)$ deriva da un modello probabilistico di riferimento.

La funzione di ripartizione empirica – per ogni valore reale x fissato – conta il numero di volte sul totale che si è verificato l'evento $(X_i \le x)$, $i = 1, 2, \ldots, n$. Pertanto, la funzione di ripartizione empirica, per ogni x fissato, è una variabile casuale e coincide con la v.c. frequenza relativa dell'evento $A_x = (X_i \le x)$.

Ebbene, per ogni x reale, la convergenza di $F_n(x)$ verso $F(x)$ consente di controllare la validità di un modello presupposto valido per i dati, il che è un momento essenziale della logica induttiva di tipo statistico.

Inoltre, poiché qualunque funzione regolare su $F_n(x)$ converge all'analoga funzione su $F(x)$, grazie al teorema di Glivenko e Cantelli, è possibile costruire su dati empirici valutazioni accurate per indicatori teorici: il che rende legittimo il fatto che si opera su un sotto-insieme dei dati *come se* fosse la popolazione di riferimento.

■ ***Teorema di Glivenko e Cantelli***

Il massimo assoluto della differenza tra la funzione di ripartizione empirica e la funzione di ripartizione converge uniformemente e quasi certamente a 0.

Scriveremo:

$$Pr\left(\lim_{n\to\infty} \sup_{x\in\mathbb{R}} | F_n(x) - F(x) | = 0\right) = 1.$$

Quindi, il teorema afferma che, per ogni x reale, è certo che la massima differenza possibile tra funzione di ripartizione empirica e funzione di ripartizione tende ad annullarsi al crescere di n.

11.5 Elementi di teoria della simulazione

Una delle applicazioni più convincenti della Teoria della probabilità per finalità operative riguarda la *simulazione*, intesa come possibilità di replicare, fin che si vuole e in modo semplice ma accurato, la struttura di un modello probabilistico per studiarne le caratteristiche statiche (distribuzione, valori medi, comportamento nelle code) e/o quelle dinamiche (la previsione, il controllo, il confronto temporale).

Questi risultati sono resi possibili da due acquisizioni teoriche:

i) si possono generare *numeri pseudo-casuali* cioè delle sequenze di numeri che possono essere ritenute realizzazioni di un prefissato modello probabilistico;

ii) si possono utilizzare tali realizzazioni per analizzare la distribuzione (o qualche opportuna sintesi), il che rappresenta il *metodo Monte Carlo.*

Il primo momento viene oggi implementato in modo particolarmente efficiente mediante algoritmi che ottimizzano gli aspetti numerici e statistici dei risultati (la maggior parte dei software disponibili possiede questi programmi).

Il secondo momento concerne tecniche statistiche e metodologie di pianificazione dell'esperimento di simulazione che rispetti le proprietà formali dei modelli e sfrutti i teoremi asintotici della teoria delle probabilità (alcuni dei quali discussi in questo capitolo).

In effetti, la teoria della simulazione statistica è piuttosto complessa per cui ci limitiamo qui ad un richiamo alle sue linee generali: di fronte ad una determinata prova, si formula un apposito modello probabilistico, si individuano i parametri di riferimento, si generano i numeri pseudo-casuali idonei per tali schemi, si specificano le quantità di cui interessa la distribuzione asintotica e, sulla base dei risultati della simulazione, si perviene a conclusioni generalizzabili.

Quindi, può essere difficile approfondire tali aspetti prima di affrontare le tematiche dell'Inferenza statistica (di cui ai Capitoli successivi); tuttavia, qui ci limitiamo ad un primo aspetto di tale problematica per anticipare l'importanza di questo approccio, nella misura in cui esso si ricollega immediatamente alla Teoria della probabilità.

- Per sperimentare in concreto schemi probabilistici astratti occorre disporre di strumenti che producano numeri casuali che siano coerenti – secondo criteri da specificare – con le realizzazioni di una v.c. prefissata. In casi molto semplici e relativi a v.c. discrete, una soluzione è immediata, anche se poco pratica.
- Per esempio, se si desiderano realizzazioni di una v.c. $X \sim Ud(6)$, la cosa più semplice da fare è quella di lanciare ripetutamente un dado ben equilibrato annotando il numero dei punti che, nei vari lanci, si presentano sulle facce. Invece, per una v.c. Uniforme continua, si può utilizzare il meccanismo dell'ago ruotante (paragrafo 10.2) segnando accuratamente l'angolo che si verifica nell'intervallo $[0, 360]$, perché esso è una realizzazione della v.c. $X \sim U(0, 360)$, e che poi si può riferire alla realizzazione di una generica v.c. $Y \sim U(a, b)$ mediante la relazione $Y = a + X(b - a)/360$.
- Tuttavia, ogni generazione meccanica di numeri casuali presenta come difetto fondamentale quello dei tempi e dei costi assieme all'esigenza di possedere oggetti fisici (monete, dadi, palline, ruote, etc.) dalla cui uniformità, regolarità e precisione dipende l'esito dei risultati.
- È invece molto più efficiente affidarsi ad un elaboratore per la generazione di numeri che, al fine di distinguerli da quelli veramente casuali, vengono definiti *numeri pseudo-casuali.* Tali metodi richiedono accorgimenti anche molto sofisticati sui quali non possiamo soffermarci.
- Si osservi, comunque, che essi sono *numeri pseudo-casuali*, e non casuali, perché le realizzazioni sono comunque ottenute con meccanismi ripetibili in modo deterministico, il che è concettualmente differente dalle prove che sono generate da meccanismi reali ad esito incerto. Per esempio, i primi numeri estratti al Lotto su una ruota prefissata negli ultimi 100 anni sono dei *numeri casuali* perché – quasi certamente – ripetendo le estrazioni non si verificheranno nella medesima sequenza. Invece, i numeri generati da un PC imponendo probabilità $1/90$ agli interi compresi tra 1 a 90 sono *numeri pseudo-casuali* perché (specificando il «seme»=*seed* iniziale dell'algoritmo) l'utente può replicarli identicamente infinite volte. Agli effetti pratici, quando l'utente non impone un seme specifico per la generazione degli stessi numeri pseudo-casuali, i software generano sequenze differenti perché il seme è collegato all'orologio della macchina e, quindi, cambia automaticamente.

Il secondo aspetto della simulazione riguarda la utilizzazione statistica dei numeri pseudo-casuali per la soluzione di un problema reale, e ciò avviene mediante il cosiddetto *metodo Monte Carlo*, che è utile in tutte quelle situazioni nelle quali la determinazione di funzioni analiticamente ben definite è difficile. Per questo, il metodo rappresenta uno strumento agevole e potente per applicare la Teoria della probabilità nella risoluzione di problemi che, altrimenti, sarebbero di lunga, difficile o impossibile soluzione.

A titolo di esemplificazione, presentiamo ora un caso reale nel quale si determina mediante simulazione la distribuzione di probabilità relativa ad un esperimento per il quale il calcolo esplicito delle probabilità sembra essere proibitivo.

Esempio 11.2

La prova consiste nello sfogliare consecutivamente le 40 carte di un mazzo di carte napoletane, accuratamente mescolate. Diciamo che si è verificato l'evento E se il numero che compare sulla carta i-esima è inferiore di una unità a quello che compare sulla carta $(i+1)$-esima, per $i = 1, 2, \ldots, 39$. In altri termini, l'evento E si verifica se e solo se – a prescindere dal «seme» – due carte consecutive presentano una delle seguenti coppie di numeri:

$$(1,2) \quad (2,3) \quad (3,4) \quad (4,5) \quad (5,6) \quad (6,7) \quad (7,8) \quad (8,9), \quad (9,10).$$

Per esempio, se escono consecutivamente {2 di Coppe, 3 di Bastone} diremo che si è verificato l'evento E; se si verifica consecutivamente la sequenza {10 di Spada, Asso di Denaro, 2 di Spada, 3 di Denaro} diremo che l'evento E si è verificato 2 volte, cioè tra la 2^a e la 3^a e tra la 3^a e la 4^a carta; e così via.

Si osservi che sfogliare consecutivamente le 40 carte equivale ad estrarre *senza ripetizione* 1 carta dal mazzo per 40 volte; dovendo esaminare i numeri consecutivi sulle carte è molto più comodo sfogliare il mazzo, dopo averlo mescolato inizialmente con molta cura.

Su tale prova definiamo la v.c. X = «numero di volte che l'evento E si verifica». Evidentemente, X è una v.c. discreta con supporto $\{0, 1, 2, \ldots, 36\}$ come è agevole convincersi. Infatti, l'evento $(X = 36)$ si verifica se e solo se le 40 carte escono ordinate dall'1 al 10 per 4 volte (per qualsiasi permutazione dei quattro semi). Per contro, l'evento $(X = 0)$ si verifica se l'evento E non si verifica mai.

Si desidera conoscere la distribuzione di probabilità di X per impiantare un gioco su tale prova e rispondere ad alcune domande:

i) qual è la probabilità che $(X = 0)$? che si verifichi $(X \geq 10)$? che si verifichi $(3 < X \leq 6)$?

ii) qual è il valor medio della v.c. X, il che è un elemento determinante per stabilire la posta in un gioco del genere?

iii) qual è la mediana di X cioè per quale valore Me è indifferente scommettere sull'evento $(X \leq Me)$ oppure sull'evento $(X \geq Me)$?

Per questo problema, la teoria della probabilità indica che, essendo ciascuna sequenza di 40 carte un evento elementare di una partizione equiprobabile dello spazio campione, basta calcolare il numero delle sequenze che generano l'evento $(X = x)$, per $x = 0, 1, 2, \ldots, 36$ e rapportarlo al numero totale delle sequenze possibili, pari a: $40!/(4!)^{10} \simeq 1.28686 \times 10^{34}$. Per rendersi conto della difficoltà di risolvere il problema in tempi accettabili, basti pensare che se si possedesse un super-computer capace di generare 1 miliardo di tali sequenze in appena 1 miliardesimo di secondo, il tempo occorrente per esaminarle tutte è dell'ordine di 10 milioni di miliardi di anni ...

Per contro, si può simulare il gioco un numero grandissimo M di volte – mediante un software appropriato – e contare quante volte si verifica l'evento $(X = x)$, $x = 0, 1, 2, \ldots, 36$, rapportando poi tale frequenza assoluta al numero totale degli esperimenti. Sulla base delle leggi dei grandi numeri (paragrafo 11.2), *la frequenza relativa di tali risultati tenderà alla probabilità dei corrispondenti eventi*, con una accuratezza tanto più elevata quanto più sarà elevato M.

Si può constatare che, in tale problema, la distribuzione si stabilizza già per $M = 1000$ prove, ed è notevolmente accurata con 500.000 simulazioni; tuttavia, nella successiva tabella riportiamo quella ottenuta dopo 10 milioni di simulazioni.

Modalità	Frequenze assolute	Frequenze relative	Funzione di ripartizione
0	232.266	0.0232266	0.0232266
1	912.691	0.0912691	0.1144957
2	1.752.779	0.1752779	0.2897736
3	2.195.584	0.2195584	0.5093320
4	2.006.869	0.2006869	0.7100189
5	1.432.132	0.1432132	0.8532321
6	826.357	0.0826357	0.9358678
7	398.457	0.0398457	0.9757135
8	161.727	0.0161727	0.9918862
9	57.475	0.0057475	0.9976337
10	17.564	0.0017564	0.9993901
11	4.652	0.0004652	0.9998553
12	1.170	0.0001170	0.9999723
13	241	0.0000241	0.9999964
14	32	0.0000032	0.9999996
15	3	0.0000003	0.9999999
16	0	0.0000000	0.9999999
17	0	0.0000000	0.9999999
18	1	0.0000001	1.0000000
Totali	10.000.000	1.0000000	==========

Si osservi che i valori per $x = 19, 20, \ldots, 36$ (che sono possibili perché fanno parte del supporto della v.c. X avendo una probabilità di verificarsi maggiore di 0) non si sono mai verificati nell'esperimento di simulazione; anzi, 1 sola volta su 10 milioni di esperimenti si è verificato un numero di carte consecutive superiore a 15.

Dalla tabella, possiamo inferire la corrispondente distribuzione di probabilità (considerando le frequenze relative della tabella *come se* fossero le probabilità della v.c. X prima definita), il cui grafico è riportato nella Figura 11.2.

Ora, utilizzando tali «probabilità simulate» (cioè le frequenze relative ottenute dalla simulazione) come se fossero le probabilità calcolate esattamente, si ottengono i seguenti *momenti caratteristici empirici*:

$$\mathbb{E}(X) = 3.59961;\ Var(X) = 3.28406;\ Asym(X) = 0.45205;\ Kurt(X) = 3.15094.$$

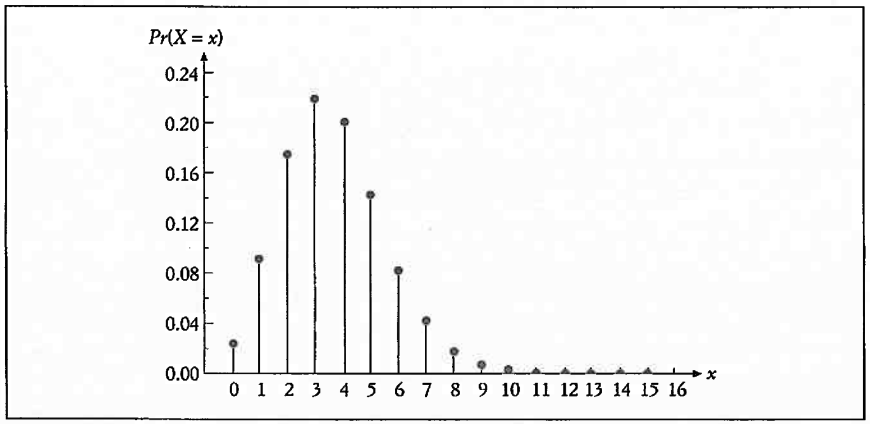

FIGURA 11.2. Distribuzione di probabilit della variabile casuale discreta dell'esempio 11.2.

Da tali indicatori (e dalla figura) emerge che la v.c. X presenta una asimmetria positiva con una forma abbastanza simile a quella di una v.c. Normale, pur se limitata ad un supporto discreto e finito. La moda e la mediana coincidono perché $Mo = Me = 3$; tuttavia, la frequenza relativa della modalità $(x = 4)$ è risultata assai prossima alla frequenza massima e la funzione di ripartizione empirica nel punto $(x = 3)$ vale circa 0.509. Ciò implica che, in un esperimento concreto, accadrà spesso che si osserveranno moda e mediana empiriche che saranno anche uguali a 4, anziché a quelle teoriche di 3.

Ora, è possibile valutare con sufficiente approssimazione le probabilità prima richieste. Esse risultano essere, rispettivamente:

$$\begin{aligned} P_r(X = 0) &\simeq 0.02322; \\ P_r(X \geq 10) = 1 - P_r(X \leq 9) &\simeq 0.00237; \\ P_r(3 < X \leq 6) &\simeq 0.42654. \end{aligned}$$

dove abbiamo utilizzato il simbolo $\simeq$ per ricordare che tali probabilità sono una valutazione approssimata (seppur affidabile) delle probabilità teoriche degli eventi corrispondenti.

Alla fine, la distribuzione simulata può essere utilizzata, per assumere qualsiasi decisione su tale esperimento, *come se* fosse la distribuzione teorica di probabilità della v.c. X, con l'indubbio vantaggio che l'intera distribuzione della tabella è stata ottenuta in tempi estremamente ragionevoli (qualche ora, su un PC portatile).

Tuttavia, non va sottovalutata una circostanza molto concreta: utilizzare questa distribuzione per scommettere sul serio, sfogliando mazzi di carte veri (e non simulati mediante il PC), può creare qualche sorpresa se non si ha l'accortezza di mescolare le carte accuratamente (molto accuratamente!), prima di procedere ad una prova successiva. Infatti, l'esperienza ha dimostrato che le tecniche usuali per mescolare un mazzo di carte da gioco non conferiscono ad ogni carta la stessa probabilità di essere in una qualsiasi posizione perché, generalmente, tali operazioni manuali sono svolte in modo affrettato.

□

Ambiente R

Teoremi asintotici

R11.1 Teorema Limite Centrale

Per sperimentare empiricamente tale teorema, generiamo n numeri pseudo-casuali da una v.c. molto differente dalla v.c. Normale e mostriamo che la somma (o la media) di tali numeri tende a distribuirsi come una v.c. Normale, al crescere di n.

A tal fine, utilizzeremo la v.c. Esponenziale negativa la cui funzione di densità è:

$$f(x) = \begin{cases} \theta\, e^{-\theta\, x}, & \text{per } x > 0; \\ 0, & \text{altrove.} \end{cases}$$

Tale v.c. ha valore medio $1/\theta$ e varianza $(1/\theta)^2$ e la sua funzione di densità è monotonicamente decrescente verso 0 al crescere di x. Come mostra la Figura R 11.1, il suo andamento è sostanzialmente differente dalla tipica forma simmetrica e campanulare della v.c. Normale. Eppure, come ora sperimenteremo, al crescere del numero degli addendi, la somma di v.c. pur così asimmetriche comunque converge velocemente alla v.c. Normale.

Consideriamo una m-pla di v.c $(X_1, X_2, \ldots, X_m)$, ciascuna delle quali è indipendente ed identicamente distribuita come la v.c. $X \sim En(\theta)$. La v.c. somma $S_m = \sum_j X_j$ possiede valore medio $\mathbb{E}(S_m) = m/\theta$ e varianza $Var(S_m) = m/\theta^2$; pertanto, la v.c. somma standardizzata Z_m è definita da:

$$Z_m = \frac{S_m - m/\theta}{\sqrt{m/\theta^2}} = \frac{\theta\, S_m - m}{\sqrt{m}}.$$

Essa, ovviamente, coincide con la v.c. media standardizzata delle v.c. $(X_1, X_2, \ldots, X_m)$.

Il Teorema Limite Centrale afferma che, per $m \to \infty$, la v.c. Z_m tende in distribuzione alla v.c. Normale standardizzata, cioè la distribuzione della v.c. Z_m è sempre più simile a quella della v.c. $Z \sim N(0,1)$, per $m \to \infty$.

A tal fine, per un prefissato m, simuliamo un gran numero di realizzazioni (per esempio, $nsimul = 1000$) della v.c. m-pla, cioé $(x_1, x_2, \ldots, x_m)$ e per ciascuna di esse calcoliamo la realizzazione z_m della v.c. somma standardizzata Z_m di cui otteniamo la distribuzione di frequenza. Se ripetiamo questo esperimento facendo variare $m = 2, 5, 10, 20$, vedremo il comportamento simulato della distribuzione della v.c. Z_m al crescere di m.

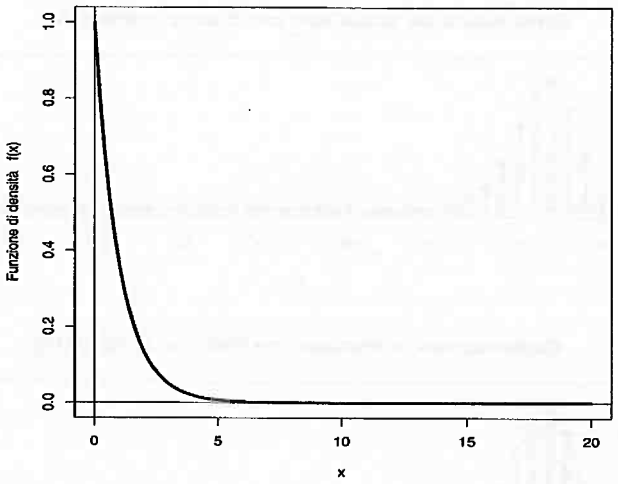

FIGURA R11.1. *Funzione di densità della v.c. Esponenziale negativa $X \sim En(\theta)$ per $\theta = 1$.*

Nell'esperimento porremo $\theta = 1$, ma il codice in R che presentiamo è utilizzabile per qualsiasi $\theta > 0$. Nel formulare le istruzioni è opportuno predisporre preliminarmente delle idonee funzioni (la cui comprensione è semplice, essendo anche commentate).

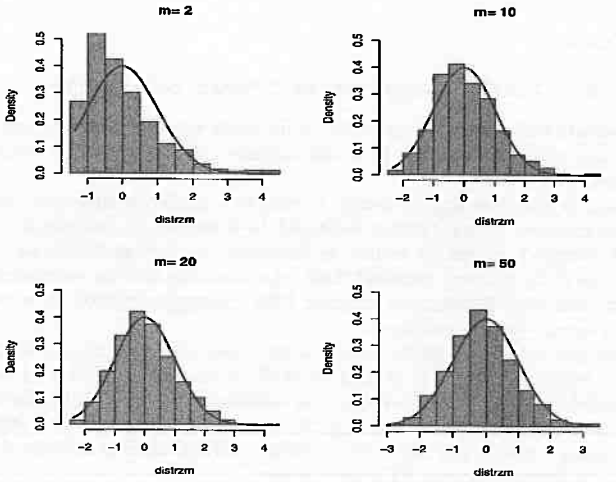

FIGURA R11.2. *Simulazione per il Teorema Limite Centrale sulla somma di v.c. Esponenziali negative con $\theta = 1$.*


```
### La funzione zetam() genera z_m
R==> zetam<-function(m,theta){
        vett=as.matrix(rexp(m,theta));
        somma=apply(vett,2,sum);
        (theta*somma-m)/sqrt(m)
}

### La funzione vettzm genera nsimul realizzazioni di z_m
R==> vettzm<-function(m,theta,nsimul){
        dzm=rep(NA,nsimul);
            for(j in 1:nsimul){
            dzm[j]=zetam(m,theta);
}
        dzm
}

### programma di simulazione e grafica
R==> nsimul=1000; theta=1;
R==> emme=c(2,10,20,50); ### Fissiamo m=2,10,20,50;

R==> par(mfrow=c(2,2))  # suddividi l'area del grafico

R==> for(nplot in 1:4){
R==> m=emme[nplot];
R==> distrzm=vettzm(m,theta,nsimul);
R==> punti=seq(min(distrzm),max(distrzm),length=100);
R==> phinorm=dnorm(punti);
R==> hist(distrzm,nclass=13,probability=TRUE,
            ylim=c(0,0.5),main="",col=gray(0.6));
R==>        title(main=paste("m=", m));
R==> lines(punti,phinorm,xlab="",ylab="",lwd=2);
}

par(mfrow=c(1,1))  # ripristina l'area del grafico
```

La Figura R 11.2 conferma il comportamento della v.c. somma standardizzata la cui distribuzione simulata converge a quella della v.c. Normale standardizzata al crescere del numero di addendi della somma, cioè per $m \to \infty$.

R11.2 Un esperimento di simulazione

L'esperimento di simulazione descritto nell'Esempio 11.2 del testo consiste nello studiare la distribuzione di probabilità della v.c. discreta X definita come il numero di volte che si verificano due valori consecutivi (a prescindere dal "seme") sfogliando di seguito e ad esaurimento un mazzo di 40 carte da gioco (contenente "figure" associate ai numeri interi da 1 a 10 e ripetute per 4 "semi" differenti). Tale v.c. assume valori in $\{0, 1, 2, \ldots, 36\}$ e lo studio effettuato consiste nel replicare un grandissimo numero di volte tale esperimento in modo da valutare la probabilità di ciascun evento tramite la sua frequenza relativa.

La programmazione è agevolata se è suddivisa in più tappe:

- costruire una funzione che mescola casualmente le carte da gioco;
- costruire una funzione consec per contare, in ogni prova, il numero di carte consecutive che si verificano;

- costruire la distribuzione delle frequenze relative dopo un numero elevato di prove.

Poiché il mazzo di carte da gioco replica i valori da 1 a 10 per i 4 semi, e l'estrazione è fatta senza ripetizione (ed a prescindere dal seme), ogni prova è generata mediante l'istruzione:

```
R==> sample(rep(1:10,4));
```

Infatti, la funzione `sample(vettore)` estrae senza ripetizione tutti gli elementi contenuti in `vettore`.

Per conteggiare i valori consecutivi, basta controllare in quali casi la differenza tra il valore della carta e quello precedente è pari a +1, cominciando dalla seconda; questo si ottiene sfruttando la funzione di R `diff` (che genera la differenza tra valori consecutivi del vettore: è stata già introdotta nel paragrafo 2.12) all'interno della seguente funzione:

```
R==> consec<-function(vettore){sum(diff(vettore)==1)};
```

Infine, la distribuzione delle frequenze relative delle osservazioni contenute in `vettore`, i cui elementi possono assumere valori appartenenti a $\{0, 1, 2, \ldots, 36\}$, si ottiene con la funzione:

```
R==> tabulate(vettore+1,nbins=36+1)/length(vettore);
```

Ciò è necessario perché la funzione `tabulate` conta le osservazioni comprese tra 1 e `nbins`, mentre nel nostro caso occorre contare quelle tra 0 e 36.

È ora più agevole sviluppare l'intero esperimento di simulazione, ed ottenere la rappresentazione della parte superiore della Figura R 11.3, mediante le seguenti istruzioni:

```
### Esperimento di simulazione replicato 1 milione di volte
R==> nsimul=1000000; quanti=1:nsimul;
R==> carte=matrix(rep(1:10,4),nrow=40,ncol=nsimul);
R==> carshuffled=apply(carte,2,sample);
R==> conteggio=apply(carshuffled,2,consec);
R==> freqrel=tabulate(conteggio+1,nbins=36+1)/nsimul;

### Rappresentazioni grafiche
R==> par(mfrow=c(2,1))  # suddividi l'area del grafico

### Distribuzione delle frequenze relative osservate
R==> plot(0:36,freqrel,type="h",lwd=2,xlab="",ylab="",
       ylim=c(0,0.25));
R==> points(0:36,freqrel,pch=19,,xlab="",ylab="");
R==> title(main="Esperimento del gioco 'Numero di carte
       consecutive'",
       cex.main=0.6,xlab="x",ylab="frequenze relative");
R==> abline(h=0);

### Confronto con le probabilità della v.c. Bin(36,1/10)
R==> plot(0:36,freqrel,type="h",lwd=4,col=gray(0.6),,
       ylim=c(0,0.25)
       xlab="x",ylab="Pr(X=x)");
R==> title(main="Confronto con la distribuzione della v.c.
       Bin(36,1/10)",
       cex.main=0.6);
R==> lines(0:36,dbinom(0:36,36,1/10),type="b",
       pch=19,cex=1,col="blue");
```


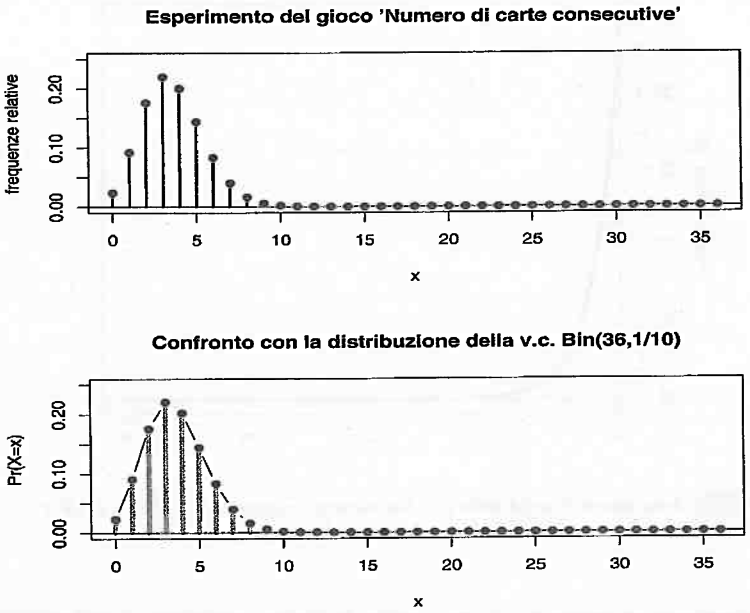

FIGURA R11.3. *Esperimento di simulazione e confronto con la v.c.* $X \sim Bin(36, \frac{1}{10})$.

```
R==> abline(h=0);

R==> par(mfrow=c(1,1))  # ripristina l'area del grafico
```

Le istruzioni riportate utilizzano per due volte la funzione `apply(obj,cod,funz)` la quale applica rispetto alle righe (se `codice=1`) o alle colonne (se `codice=2`) la funzione `funz` applicata all'oggetto `obj`.

Nello specifico, la funzione `apply(carte,2,sample)` applica a ciascuna colonna (perché `cod=2`) della matrice `carte` (prima definita) la funzione di campionamento senza ripetizione (cioè `sample`). In modo simile, la funzione `apply(carshuffled,2,consec)` applica alle colonne della matrice `carshuffled` (che contiene *nsimul* estrazioni senza ripetizione dell'esperimento) la funzione `consec` (che conteggia quante carte consecutive sono presenti in ciascun vettore colonna).

Il codice in R qui riportato replica i risultati del testo (cioè un milione di simulazioni) in poco più di 1 minuto su un PC dotato di 3GB di memoria RAM e si segnala per la velocità delle elaborazioni ottenute senza mai utilizzare un *loop*. Variando l'istruzione *nsimul* = ... si possono generare altri esempi; in ogni caso, la forma della distribuzione si stabilizza abbastanza presto e, per avere informazioni sull'ordine di grandezza delle singole probabilità, sono sufficienti appena 1000 simulazioni.

Poiché la seconda parte del codice sopra riportato genera anche un adattamento con la v.c. Binomiale, aggiungiamo pochi commenti, anche sulla base della Figura R 11.3.

- La forma della distribuzione è fortemente asimmetrica positiva e produce probabilità di fatto nulle per valori di $x > 15$. Quindi, un numero di carte consecutive

maggiore di 15 può ritenersi un evento estremamente raro e quasi impossibile da osservare.

- È sorprendente l'adattamento quasi perfetto tra la distribuzione di tale v.c. X e quella di una v.c. Y Binomiale che assume valori tra 0 e 36 con probabilità pari a 1/10; infatti, la massima differenza assoluta che si osserva tra le due distribuzioni è 0.0016241. Si osservi che la v.c. $Y \sim Bin(36, 1/10)$ è idonea per rappresentare il numero di "successi" in una sequenza di 36 prove *indipendenti* nelle quali la probabilità del "successo" è costante e pari a 1/10.
- A conferma dell'adattamento osservato, si consideri che una v.c. $Y \sim Bin(36, 1/10)$ possiede valore medio, varianza, asimmetria e curtosi pari a 3.6, 3.24, 0.444 e 3.142, rispettivamente, mentre le corrispondenti quantità (empiriche) calcolate sulla base della distribuzione simulata sono 3.6006, 3.2922, 0.46126 e 3.1749, rispettivamente. Tali risultati sono coerenti con i valori 3.59961, 3.28406, 0.45205 e 3.15094, rispettivamente, ottenuti mediante la simulazione di 10 milioni di prove presentata nel paragrafo 11.5 del testo.

Parte terza

Inferenza statistica